Deflating Representation
Computation, Structure, and Content

Coelho Mollo, Dimitri

Awarding institution:
King's College London

Deflating Representation
Computation, Structure, and Content

Dimitri Coelho Mollo

Submitted for the degree of PhD in Philosophy
at King's College London, Department of Philosophy;
and Humboldt-Universität zu Berlin, Philosophische Fakultät I,
Dekanin Prof. Dr. Gabriele Metzler

Examination passed on 11.09.2017
Supervisors: Prof. Nicholas Shea, Prof. Dr. Michael Pauen, Prof. David Papineau
Examiners: Prof. Michael Rescorla; Dr. Mark Sprevak; Prof. Dr. Karl-Georg Niebergall

Abstract

The present work focuses on the notions of representation and computation, and the explanatory role they play in the cognitive sciences. I put forward a deflationary view of representational content, and argue that explanatory internal states in the cognitive sciences are primarily individuated by their computational structure, rather than by content. In Part I, I survey the mainstream accounts of representation and content present in the philosophical literature: functional role semantics, informational semantics, teleosemantics, and structural representation. I also briefly examine some of the crucial issues that any satisfactory theory of content has to tackle, with special attention to the problem of indeterminacy of content.

I present and develop, in Part II, a version of the mechanistic view of concrete computation able to account for how cognitive systems compute, and for how to individuate their computational structures. The account avoids pancomputationalism and triviality of computation, yielding a robust, objective theory of computation in physical systems.

With this mechanistic view of concrete computation in hand, in Part III I present my deflationary approach to representation, which shifts much of the explanatory burden in making sense of cognition onto computational structures. I examine interpretational semantics as a promising precursor of my view. I propose several modifications to interpretational semantics, producing a theory close to structural representation, but with marked deflationary leanings. On that basis, two deflationary paths are examined: content pragmatism, and mild realism about content. I provide reasons to prefer the latter approach, though I take both paths to be promising. The resulting deflated notion of representation, wedded to a solid notion of computational structure, is advantageous insofar as it dissolves metaphysical puzzles related to content-fixation and indeterminacy, while preserving a notion of representation robust enough to play an important explanatory role in the contemporary study of cognition.

Zusammenfassung

Schwerpunkte dieser Dissertation sind die Begriffe von Repräsentation und Komputation, insbesondere in Bezug auf ihre Erklärungsrolle in den Kognitionswissenschaften. Ich entwickle eine deflationäre Theorie bezüglich des Gehalts von mentalen Repräsentationen und spreche mich dafür aus, dass die inneren Zustände, die eine Erklärungsrolle in den Kognitionswissenschaften spielen, eher durch ihre komputationale Struktur bestimmt werden als durch ihren Gehalt.

Im ersten Teil meiner Dissertation gebe ich eine Zusammenfassung der bekanntesten Theorien des repräsentationalen Gehalts in der philosophischen Fachliteratur. Das sind die inferentielle Semantik, die Informationssemantik, die Teleosemantik, und die strukturale Repräsentationstheorie. Außerdem analysiere ich einige der wichtigsten Probleme, auf die jede Theorie des repräsentationalen Gehalts eingehen muss, mit besonderem Fokus auf dem Problem der Unbestimmtheit des Gehalts.

Im zweiten Teil präsentiere und entwickle ich eine Variante einer mechanistischen Theorie der konkreten Komputation, die in der Lage ist zu erklären, wie Komputation in kognitiven Systemen möglich ist und wodurch komputationale Strukturen charakterisiert werden. Diese Theorie vermeidet das Risiko eines Pankomputationalismus und einer Trivialisierung der Komputation. Zudem liefert sie eine robuste, objektive Erklärung der Komputation physikalischer Systeme.

Anhand dieser mechanistischen Theorie der konkreten Komputation präsentiere ich im dritten Teil meine deflationäre Theorie der Repräsentation, in der die in den Kognitionswissenschaften benötigte Erklärungsleistung von komputationalen Strukturen übernommen wird. Ich beginne mit einer Analyse der interpretationalen Semantik, die ein vielversprechender Vorgänger meiner Theorie ist. Daraufhin schlage ich diverse Änderungen an der interpretationalen Semantik vor, wodurch eine Theorie entsteht, die ähnlich der strukturalen Repräsentationstheorie ist, aber eine starke deflationären Neigung hat.

Auf diesen Grundlagen werden zwei deflationäre Vorschläge analysiert: Pragmatismus und milder Realismus über den repräsentationalen Gehalt. Ich gebe Gründe dafür an, dass der letztgenannte Vorschlag zu bevorzugen ist, obwohl ich glaube, dass beide Vorschläge aussichtsreich sind. Das resultierende deflationäre Konzept der Repräsentation, kombiniert mit einem robusten Konzept der komputationalen Struktur, hat den Vorteil, dass es metaphysische Probleme in Bezug auf die Festlegung und Unbestimmtheit des repräsentationalen Gehalts löst. Auf diese Weise wird ein robustes Konzept der Repräsentation bewahrt, das ausreicht, eine wichtige Erklärungsrolle in den Kognitions-

wissenschaften zu spielen.

Contents

List of Tables

Preface

Deflating Representation: An Overview

Cognition is to be studied and understood to a significant extent by means of the notions of representation and computation. Or, at least, this is the insight that gave birth to Cognitive Science, a research field that has been thriving and expanding for the past 60 or so years. Areas of psychology, linguistics, artificial intelligence, philosophy, neuroscience, sociology, and anthropology have embraced the idea, coming to form the mosaic unity of the study of cognition, to borrow a phrase from Craver (2007). Alternative proposals, ones that do not see cognition as a representational, computational phenomenon, have been surely put forward in the past decades, but they have largely failed to overthrow the representational-computational framework, which dominates research in the field to this day. Cognitive psychology, cognitive neuroscience, philosophy of cognitive science, psycholinguistics, computational neuroscience, to name a few: Cognitive Science has directed and spawned several sub-fields, which, despite occasional hiccups and threats of crisis (*cf.* the recent replication crisis in psychology), have been moving forward and enriching our understanding of cognitive phenomena.

Philosophers, fittingly, have not failed to worry about the conceptual underpinnings of the whole project. Representation and computation are terms that have become part of everyday discourse (though the latter much more recently than the former), and despite a certain intuitive grasp of what those concepts capture, careful perusal reveals difficulties that undermine our certainty — and introduce cracks in the foundations of Cognitive Science. What states and processes in the world can be fruitfully considered to be representations? How can physical entities and processes have semantic content, be about other entities and processes? What is it for a physical system to compute, and which systems, if any, do so? Are representations and computations facts of the matter about cognitive systems, or are they just useful concepts that we employ to allow us better to understand cognitive phenomena? These are all pressing questions, and which prove recalcitrant to straightforward solutions.

Especially from the 1980s on, philosophers have taken up the task of trying to secure the sustaining pillars of Cognitive Science by offering answers to these questions (and some more). Much effort went into projects that attempted to make the notions of representation and computation precise and scientifically acceptable. The bulk of this project went (and goes) to trying to naturalise representation and computation — that is to say, to explain or reduce those notions in terms of entities and processes that are

the bread and butter of more basic sciences, such as physics, chemistry, and biology. This naturalisation project, especially when applied to representation, had initially seemed "both fundamental and solvable"[1], but soon found itself instead having to wade sluggishly through murky, hurdle-ridden waters, and produced, after four decades of focused toil, only partial satisfaction to its tribulations. Rather than fixing the cracks in the foundations of Cognitive Science, philosophical work has revealed how deep they actually go. Hope is not lost though, and philosophical work on these issues is ongoing, albeit, to use again a phrase from Godfrey-Smith, it seems to have "lost momentum" with the turn of the century.

My project takes the cue from this nagging feeling of dismay about the prospects of success of the philosophical naturalisation project, in its attempt to provide the bases for the multifarious edifice of Cognitive Science. An important part of my claim echoes Godfrey-Smith's (2006) growing suspicion that "we have been looking for the wrong kind of theory" all along. Though my positive approach differs considerably from his, it shares its guarded optimism. It is not that the cracks in the foundations cannot be fixed, it is just that we have been using the wrong materials to do it. In particular, I will suggest that the notion of representation can be deflated — its metaphysical and explanatory importance downsized — in at least two different ways, whilst fulfilling its foundational role in Cognitive Science. The success of such deflation of representation will pass through the proposal of a robust notion of computation, able to take an important share of the explanatory burden on its shoulders.

This work is divided into three parts.

The aim of Part I is to introduce the debate over the role of representation in cognition, focusing particularly on attempts at naturalising the notion. After clarifying the explanatory roles that the notions of representation and computation play in Cognitive Science, Chapter 1 gives a quick overview of the most influential theories that have been proposed to try and naturalise representation. I briefly examine four views: functional role semantics, causal-informational semantics, teleoinformational semantics, and teleosemantics. I also present seven traditional objections that have been moved against those views. These point to problems that a satisfying theory of representation should try and avoid. I focus especially on problems relating to indeterminacy of content. One central *desideratum* for theories of representation is to provide conditions on content-fixation that yield determinate or fairly determinate representational contents. Indeterminacy of content risks jeopardising the explanatory purchase of representation, as well as the possibility of misrepresentation. Given the limited objectives of this chapter, I will not but offer a rough and ready treatment of those views and issues.

In Chapter 2, I analyse a theory of representation that has a venerable history, but which has partially lost its place under the philosophical spotlight in the second half of the last century: structural representation. This theory is generally taken to be a sophisticated version of resemblance theories of representation, according to which representations represent by means of resembling what is represented. A theory on those

[1] The phrase is Godfrey-Smith's (2006).

lines, I will later argue in Part III, can provide, suitably modified in its commitments, a useful basis for a deflationary view of representation. I show that, as it stands, structural representation has its own indeterminacy problems: the content-fixing resemblance relation it relies on is too liberal, making representational content wildly non-unique. After going through some recent attempts to further refine the theory in order to avoid that unfortunate consequence, I argue that they fall short from sufficiently curbing pernicious liberality, and moreover risk falling into the traditional objections examined in Chapter 1.

Part II may initially appear as a sudden change of subject. I leave issues about representation behind for a moment, and focus instead on the notion of concrete computation, that is to say, computation in physical systems (vs. in mathematical theory). This change of subject is only apparent: providing a robust theory of concrete computation is instrumental for my deflationary account of representation, to be presented and defended in Part III.

Chapter 3 gives a detailed overview of the most influential theories proposed to naturalise computation, and thus yield an account of concrete computation. I present and examine the features and shortcomings of the simple mapping view of computation, as well as of a pragmatic take on it; and then move on to assess Chalmers' (2011) sophisticated causal account; concluding the chapter with a treatment of the semantic view of computation. I argue that these four views display several points of dissatisfaction, and should be rejected in favour of a fifth: the mechanistic view of concrete computation.

A careful treatment of the mechanistic view of concrete computation is offered in Chapter 4. I clarify the nature of the overall mechanistic approach to scientific explanation, turning subsequently to its application to computation. I focus on perhaps the most worked-out version of the mechanistic view of computation, developed extensively by Piccinini (2007*b*, 2008*a*, 2015). Though I find much to agree with in Piccinini's treatment, I propose to amend his view in two significant ways. Firstly, I propose a better way of seeing the role played by the appeal to mechanisms in the account, thereby avoiding a dilemma that could otherwise prove fatal to the account. Secondly, I put forward a more satisfying reply to one of the most powerful objections against non-semantic views of computation: the argument from the multiplicity of computations. This amended version of the mechanistic view plays an important role in allowing me to discharge some of the metaphysical weight normally put on the notion of representation onto the robust notion of concrete computation instead. A final worry has to be put to rest, though. The mechanistic view hinges on teleological functions of a certain sort: physical computational systems are mechanisms with a specific teleological function, *i.e.* the function to perform computations. In order to earn its keep, it must be shown that suitable theories of teleological function are at hand.

Chapter 5 is meant to allay these fears regarding the mechanistic view of concrete computation by examining several extant accounts of teleological function, and showing that at least some of them are appropriate to fulfil the needs of the view. I start by making clear the crucial role that the appeal to teleological functions plays in the

mechanistic theory, and thereby the features that these should have if they are suitably to play that role. Specifically, teleological functions help make concrete computation something objective, relatively rare, normative, and compatible with our pretheoretical intuitions about computing systems. The question is then whether there are theories on the market that can buy the mechanist all these much wanted goods. I examine several theories of function, belonging to two main families: dispositional theories, which comprise the analytic account, perspectivalism, propensity theories, and goal-based theories; and selected-effects theories, which include theories that appeal to different ranges and types of selection processes. My assessment of the available options justifies hopefulness toward the mechanistic view of concrete computation. I argue that at least two types of theories — goal-based theories and broad selected-effects theories — provide notions of function capable of doing the job required by the mechanistic view. The mechanistic view of concrete computation, I argue, proves therefore to be a robust and satisfying theory — and one able to play the role that my deflationary view of representation requires. Part II closes on this rather positive note.

In Part III, the various open threads come together in the build-up toward the deflationary view of representation I propose. In particular, in this part I rely on the robust notion of concrete computation provided by the mechanistic view to individuate computational structure as one of the factors that carries the most load in explaining complex appropriate behaviour. Ascription of determinate representational content comes on top of that, and heavily depends on context — explanatory, current, and historical. The crucial property of cognitive states that plays a primary role in explanation of behaviour is then computational structure, with representational content being relegated to a secondary, albeit ineliminable role. By the lights of this deflationary approach, representational content and representational vehicles may lack many of the characteristics that they are often believed to have, such as fairly clear boundaries, stability, and repeatability. This view can be developed in (at least) two ways: pragmatism about content, and mild reductionism.

Chapter 6 presents and examines a relatively underrated theory of representation: interpretational semantics. This analysis is the first stepping stone leading to the deflationary view of representation I intend to defend. I examine two versions of the view, put forward by Cummins (1989) and Ramsey (2007). My assessment points out several shortcomings in the views as they stand, especially with respect to the weak notions of concrete computation they employ — though Ramsey's version seems to hold some promise. I also show that those views may run afoul, as the other theories examined before, of indeterminacy of content problems, leading to pernicious non-uniqueness of content.

Despite its extant shortcomings, I believe that interpretational semantics has much to recommend it. In Chapter 7 I show how, by accepting the robust notion of computation provided by the mechanistic view, interpretational semantics can lay the basis for a deflationary version of structural representation.

Chapter 8 finally brings my project to a close. It presents and defends two paths to-

ward deflationism about representation and content. The first takes its inspiration from Frances Egan's content pragmatism. I argue that my approach improves on hers insofar as it employs a better notion of computation, and eschews her appeal to mathematical content, insisting rather on the importance of computational structure. I also defend content pragmatism from recent objections, and show that they fail to undermine the view.

The second path that I develop is non-pragmatist in nature, though it also rejects the assumptions behind mainstream naturalistic theories of representation and content. I claim that content should be seen as a real pattern in nature (Dennett 1991) — a pattern of variable, diverse, and indeterminately long conjunctions of sundry factors, all brought together and rendered salient by an underlying rationale or perspective: the adaptation of organisms to their historical environments. While naturalistic, this version of the deflationary framework, in contrast with mainstream theories, denies that there is a definite set (or a limited number of sets) of conditions that bestow content on representational vehicles. The complexity and context-sensitivity of the content pattern denies such a possibility. By seeing organisms as adapted to their environments, a host of factors, both occurrent and historical, become salient as helping to explain, case-by-case, robust successful behaviour in face of changing environmental conditions. An essential condition for the content pattern to emerge is the presence of suitable computational structures in the cognitive system. Therefore, computational structure has pride of place in this version of deflationism as well. I argue that this second deflationary path has theoretical virtues that the first lacks, and thereby should be preferred.

Before bringing to a halt the long and windy route that brought me to deflationism about representation, I offer some remarks on how the outcomes of my project bear on current empirical and theoretical issues in the cognitive sciences, with special regard to the debate on cognitive ontology. I argue that recent evidence, which suggests that the brain reuses the same neural resources in cognitive tasks of very different nature, lends force to the ideas proposed here. Multi-purpose neural circuits carry out computations that can be usefully recruited and combined to give rise to the overall computational processes that enable successful behaviour in different cognitive tasks. This gives reasons to believe that the primary explanatory role in the cognitive sciences, as I urge in this work, is played by computational structures, with representational content having a secondary, and perhaps merely pragmatic function to perform.

These considerations will bring this work to its end, and hopefully will have further substantiated the plausibility, both conceptual and empirical, of the views it defends. Though I do not expect that it will convert all minds to the cause, my objectives will have been achieved if, at least, I will have shown that the deflationary approach merits attention and study, and that it can appear as an appealing candidate for providing a satisfying theory of representation and content that is useful for the advancement of the cognitive sciences.

Acknowledgements

I have had the help of many people in writing this dissertation. I am indebted to my supervisors, Nicholas Shea, Michael Pauen, David Papineau, and Matteo Mameli for precious support, and many engaging and illuminating discussions. In addition, I owe much to many philosophers and cognitive scientists that I have encountered during these years. Especial mention goes to Robin Guido Löhr, Juan Raul Loaiza, Daniel J. Cook, Astrid Schomäker, Nir Fresco, and Margherita Arcangeli. I am also indebted to my examiners, Michael Rescorla, Mark Sprevak, and Karl Georg Niebergall, for their fruitful and valuable comments.

First and foremost, I am indebted to my parents, Alexandre Miranda Mollo and Marcia Cristina Coelho Mollo, for everything.

Chapter-specific acknowledgements:

Chapter 4: I am indebted to Nir Fresco and Oron Shagrir for discussion of drafts of this chapter. Material adapted from this chapter has been published as the paper "Functional individuation, mechanistic implementation: the proper way of seeing the mechanistic view of concrete computation", *Synthese* (2017), doi: 10.1007/s11229-017-1380-5.

Chapter 7: I am indebted to Brian Ball, Frances Egan and audiences at the 3rd HaPoC Conference 2015, in Pisa, the 2nd Trends in Interdisciplinary Studies Conference 2015, in Warsaw, the NCH Mind and Brain Conference 2016, in London, the IACAP meeting 2016, in Ferrara, and the BSPS Conference 2016, in Cardiff, for valuable feedback on early versions of this chapter.

Chapter 8: I am indebted to Brian Ball, Frances Egan and audiences at the NCH Mind and Brain Conference 2016, in London, the IACAP meeting 2016, in Ferrara, the BSPS Conference 2016, in Cardiff, as well as an anonymous referee of *Topoi* for valuable feedback on early versions of parts of this chapter. Material adapted from section §8.1 has been published as the paper "Content Pragmatism Defended", *Topoi* (2017), doi:10.1007/s11245-017-9504-6.

Part I

Representation and computation in the cognitive sciences

Chapter 1

The representational, computational mind

With the advent of Cognitive Science as a research paradigm in the 1950s, the internal states of cognitive systems came to be seen as crucial elements for explaining the behaviour of organisms. Some of those internal states are taken to inform the cognitive system about the goings-on in the body and environment of organisms, therefore guiding behaviour that is appropriate to their circumstances. Such internal states may also interact with each other in sundry ways, allowing a further degree of complexity in the processes that lead to appropriate behaviour. The positing of internal states that have these properties and carry out these functions in the economy of the cognitive system requires that the having of those properties and the carrying out of those functions be explained in their turn. How can internal states 'inform' cognitive systems about what is going on outside the system itself? How can they guide behaviour? In which ways do they interact, and how are those interactions governed?

If Cognitive Science is to be a respectable scientific endeavour, all the answers to these questions must be scientifically acceptable — they can involve appeal exclusively to entities and processes that are, or are plausibly explained by means of, entities and processes already present in more basic sciences, such as biology, chemistry, and physics. Moreover, if Cognitive Science is to be informative, the explanations of cognition and behaviour it offers must themselves not ultimately involve those properties and functions that it set out to explain, on pain of vicious circularity.

Two notions have been invoked to help make more precise the vaguely expressed ideas above. Those states internal to the cognitive system that provide information[1] about the body and environment, and thus help to guide behaviour, are cognitive representations. The processes by which cognitive representations are transformed and interact with each other are computations — transitions of physical states that respect computational rules. Representation and computation lie at the foundations of Cognitive Science. They are the fundamental notions grounding the explanatory project of Cognitive Science, and are, as is to be expected, routinely used in informing empirical

[1] I am using the term 'information' in a non-technical sense.

hypotheses about cognition, in interpreting experimental data, and in putting forward theories of cognition.

I will often prefer to talk about the 'cognitive sciences', rather than about 'Cognitive Science'. By using the plural expression, I intend more fairly to capture the interdisciplinary nature of the study of cognition, and the variety of different approaches and fields of research — some of which to some extent at odds with the aims and assumptions of Cognitive Science as it was originally conceived — that have been developed to investigate cognitive phenomena in the course of the almost 70 years since the so-called Cognitive Revolution. The neurosciences, cognitive psychology, artificial intelligence, as well as areas of robotics, anthropology, linguistics, and philosophy are all concerned with the study of cognition — though often employing radically different methods — and are therefore rightly characterised as composing the cognitive sciences.

There are and there have been frameworks that purport to explain some or all cognitive phenomena without making use of the notions of representation and computation. These latter approaches, recently covered under the usefully synthetic umbrella-term 'Radical Embodiment', either intend to ground the study of cognition on notions other than representation and computation, or attempt considerably to diminish the role that those notions play in explanations of cognitive phenomena. In my terminology, frameworks such as Radical Embodiment qualify as parts of and contributors to the cognitive sciences, though they certainly would not want to run under the more exclusive banner of Cognitive Science.

At any rate, I will have little to say about anti-representational and anti-computational approaches to cognition. Friends of Radical Embodiment may be right when they claim that representation (and computation, in some understandings of the notion) may not be needed for some cognitive feats. However, many cognitive abilities, which Clark & Toribio (1994) have dubbed 'representation-hungry', call for explanations in representational and computational terms. I will accept, with most of the literature, that many interesting cognitive phenomena cannot be accounted for if not by bringing to bear the notions of representation and computation.

In order to even start assessing whether representation and computation are notions that should rightly lie at the foundations of our understanding of much of cognition, we must be clear on what those notions are, and what work they are supposed to do. While most cognitive scientists employ these notions without much forethought — as scientific posits without which their hypotheses and interpretations of empirical data would often not even make sense — philosophers have worried about the naturalistic status, and explanatory role of representation and computation. Philosophers have been concerned with providing theories of the nature of representation and computation that are able to provide notions that fulfil the explanatory role they are to play in the cognitive sciences, whilst appealing only to entities and processes that are scientifically acceptable — that are *bona fide* elements in the scientific, materialistic worldview. In brief, philosophers have been interested in naturalising representation and computation. The present work mostly accepts the framework within which those projects have been carried out, while

taking issue with some elements.

This chapter and the next will provide the lay of the land in the debate about naturalising representation, while a closer look at the naturalising computation project will have to wait for Part II, though I will provide some hints very shortly. The purpose of this somewhat rough-and-ready sketch is to provide the background against which my project is meant to stand out.

Here is how I shall proceed. In section §1.1 I briefly clarify the explanatory role of representation in the cognitive sciences. In section §1.2 I do the same for the notion of computation. After drawing some distinctions that will be useful throughout this work in section §1.3, I move on, in section §1.4, to provide a survey of the naturalistic project about representation, briefly examining the mainstream theories of representation and content on offer in the literature: functional role semantics, causal-informational semantics, teleoinformational semantics, and teleosemantics. In section §1.5 I enumerate traditional problems that theories of content have to face, with particular attention given to issues relating to indeterminacy of content. Theories of content, if they are to fulfil their aim of providing notions of representation that are scientifically acceptable and explanatorily useful, should better find ways around those problems, either by satisfactorily answering them, or keeping them from arising. My analysis of influential current theories of content will go on in the next chapter, where I present and analyse resemblance theories of content, which have fallen out of fashion despite a long history of adherents. I focus on their most promising version: structural representation.

1.1 Representation: what and why

The notion of representation in the cognitive sciences provides the tools to understand how cognitive systems can receive, store, and process information about what is going on in the body and environment. The basic idea is that representations are cognitive states which are about entities and processes in the world, in a way not too dissimilar to how (naturalistic) pictures are about what is depicted, words are about what they refer to, and maps are about the landscapes they map. The intuition is that somehow, by being about things in the world, representations stand in for what they represent (Godfrey-Smith 2006). By operating over its own representational states, cognitive systems would thereby be able to generate behaviour that is appropriate to the circumstances the organism finds itself in, and which some of its internal states are about, *i.e.* stand in for, or represent.

More precisely, and perhaps less intuitively, representations are minimally those states that bear semantic properties (Ryder 2009*a*). Representations have contents — what they are about — and those contents place conditions on how the world is or should be, if the representation is, in a loose sense, to be appropriate. The physical realisations of representational states are representational vehicles: the physical states that carry representational content. In the case of the brain, for instance, representational vehicles are, ultimately, states and processes going on in neurons and/or in populations of neurons.

Representations can be part of internal states that have a descriptive nature. When participating in such internal states, representations aim at corresponding, to some level of accuracy, to how things are, or were, or will be, in the world. In this case, it is the world that, so to speak, is dictating the terms. A belief with the content *that there is a puffin in front of me now* attempts to capture how things are in the area of the world immediately in front of the organism entertaining the belief. The belief has thereby truth conditions: it is true in case its content accurately describes the state of the world in front of the belief-entertainer, and false otherwise (for instance, were I to entertain that belief now, it would be false). Representations may also be part of internal states that have a directive nature. When participating in such internal states, representations pose constraints on how the world is to be so as to satisfy their contents. Desires are an example of directive internal states. The content of my desire *that I eat some ice-cream* is satisfied only in case the world is transformed in such a way as to bring it about that I eat some ice-cream. The desire cannot be true or false, but it can be satisfied (in case I eat some ice-cream) or unsatisfied (in the unfortunate case I do not).

Representational contents are semantic properties of states, understood in their turn as correctness conditions; be they truth conditions (as in beliefs), satisfaction conditions (as in desires), or more generally, accuracy conditions. Intuitively, these correctness conditions 'say', or 'mean', how things other than the representational state itself are, were, will or should be — in this sense, they are about those things: they have *aboutness*, as this property of representations is often called.

Beliefs and desires are relatively complicated cognitive states that belong to the category of propositional attitudes. Simpler cognitive states might though share some of their properties. In this work I will have little to nothing to say about propositional attitudes, as my focus is on the notion of representation relevant for simpler cognitive states, short of beliefs, desires, and other propositional attitudes. Propositional attitudes and the kinds of representations that are involved in them have largely monopolised the attention of philosophers of mind.

My aim in the present work is to investigate notions of representation that are useful for the cognitive sciences — my project therefore belongs to the philosophy of the cognitive sciences more than to the philosophy of mind. The kinds of internal states that much of the cognitive sciences are concerned with fall short of full-fledged beliefs and desires. They are rather states that play a role in the nuts and bolts of cognition, informing the workings of subsystems of the cognitive system, such as sensorimotor processing, spatial navigation, linguistic processing, and automatic recognition and categorisation. In contrast to propositional attitudes, such cognitive states do not belong to the whole person, and are not accessible to consciousness: they are subpersonal states.

Cognitive scientists routinely use the notion of representation (and of computation) in their hypotheses and explanations about how such subpersonal processes work. We see claims about cells in cortex area V1 representing edges in the visual scene, cells in the entorhinal cortex-hippocampus system representing spatial locations, regions of the brain processing parse trees and storing word meaning. I will be concerned in the

foregoing with notions of representation that can be useful for those explanatory purposes, though some issues that have exercised philosophers interested in propositional attitudes will surface here and there, since some of the puzzles regarding representation and computation are to be found in the treatment of personal and subpersonal states alike.

'Representation' (and 'computation') are terms that live many different lives, playing different explanatory roles and capturing different phenomena in the various fields of research in which they are employed. In the sciences of the mind, I take it that it is of vital importance to distinguish those two different scientific and philosophical questions. To avoid confusion, I use the expression 'representational content' to refer to the contents of subpersonal states, to which I normally refer as 'cognitive states', while I reserve the expression 'intentional content' to capture the content of personal states, such as propositional attitudes, which I call 'intentional states'[2]. Though it is often assumed that theories of intentional states and intentional content will also apply with some tweaking to cognitive states and representational content, I will try and keep the two issues separate. Given that the *explananda* are different, we should expect the *explanantia* to be distinct, or, at any rate, we should not assume that they are not. My interests in this work will be limited to cognitive states and representational content, and my treatment of these issues may or may not shed light on intentional states and intentional content.

Four (or five) questions about the nature of mental representation (in general, *i.e.* cognitive and intentional) drive much of the contemporary philosophical work on this topic: a) are representations objective features of the world; and what is their explanatory status in the cognitive sciences?; b) what makes so that some states function as representations (what Ramsey 2007 has called 'the job description challenge')?; c) how do representations get their contents?; d) what format do representations have (*e.g.* language-like, pictorial, local, distributed, etc.)?

Though distinct, these questions are closely related. Part of what makes a state into a representation is the fact that it has content — thereby answers to c) bear on answers to b). It is to be expected, in addition, that answers to b) and d) will strongly bear on each other: the format of representations is relevant for understanding how they play a representational role in the cognitive system. The format of representations may also inform the way representations get their contents: for instance, by having a pictorial format, representations may have their contents fixed by what they pictorially resemble. Finally, answers to a) inform answers to b) and c). Depending on the ontological status that representations have, as well as their explanatory role in the cognitive sciences, the acceptable factors contributing to making them into representations, and bestowing content on them, vary.

My aim in this work is to explore issues regarding especially a) and c): the ontological status of representations, their explanatory role, and the ways they get their contents. I will not focus directly on b) and d), though given the close relationships between those

[2] A similar distinction can be found in Cummins (1989), Coelho Mollo (2015).

four questions, my treatment of the former bear to some extent on possible answers to the latter. A further issue that will play a central role in the foregoing is the question of whether explanations in the cognitive sciences should primarily involve appeal to representational contents, as is generally believed, or to some other property. This discussion will have to wait for Part III.

I will more carefully present and examine approaches to the ontological and explanatory status of representations, and the way they get their contents in section §1.3 and section §1.4. But before moving to that, let me first spend a few anticipatory words on the nature and role of computation in the cognitive sciences.

1.2 Computation: what and why

In the cognitive sciences, the notion of representation often comes hand-in-hand with that of computation. The reasons for this pairing are not difficult to understand: representations give us only part of the story about how cognitive processes are able to lead to appropriate behaviour. They help explain how come states internal to the cognitive system are able to be informative of circumstances outside. However, we still need tools to understand how representations get transformed, how they lead to other representations in appropriate ways, and how they influence the organisms' effectors so as to lead to adequate behaviour. If behaviour is to be successful, transitions of representational states in the cognitive system must be regimented: they must make so that the interaction and sequentiality of tokened internal states follow some rule — be it of association, probability, plausibility, or rationality — that respects the representational content of those states and is appropriate to the behavioural task at hand.

Computation comes into the picture to play two fundamental roles: help explain how representational states transition and interact in a regimented way with other representational and non-representational internal states — whatever that regimentation might be, *i.e.* whatever the rule governing the transitions — and help explain how transitions and interactions between internal states take place in a way that respects their representational content — semantic properties not being themselves candidates for causal efficacy. These two roles are closely related. Let us briefly tackle how computation is supposed to carry them out.

Computation is roughly a process that leads from inputs to outputs according to well-defined transition rules. Inputs to a computational system are transformed according to rules sensitive to some of their properties, as well as, in some cases, to the internal state of the system itself, yielding outputs that depend on the inputs, internal states, and the rules applied. A pocket calculator, for instance, takes inputs in the form of button presses and, depending on which buttons are pressed, leads to an output in the form of a visual pattern on a screen. This is only one, and one of the least interesting, descriptions of what the pocket calculator does. At a representational level of description, what the device does is to take numbers and arithmetical operations as input (*e.g.* 3+7), transform those numerical values according to the arithmetical operation, and thus generate the appropriate output, the result of the calculation. What makes it so

that the calculator generates the correct result given the inputs is the fact that it transforms the representational vehicles (*e.g.* the patterns of electric activation that stand for the numbers 3 and 7) in a way that respects the rules of arithmetic, thus generating as output other representational vehicles that represent the result of the calculation.

The pocket calculator, as any other computational system, is not directly sensitive, in its operation, to the semantic properties, the content, of the internal states it traffics in[3]. The state-transition rules it follows are sensitive, rather, to some specific types of physical properties of those internal states, which constitute the syntax of the system (in opposition to its semantics). What is most interesting for our purposes, though, is that by organising the syntax of a computational system in an appropriate way, the state-transitions can be such that they mirror what goes on in another domain — the computational transitions are such that they correspond to the transformations of elements in the other domain. In case the other domain is a set of true propositions about the world, for instance, the computational transitions will preserve truth, leading to other true propositions about the world. Thereby internal states and processes become interpretable as standing in for entities and processes in the world (inclusive of the abstract 'world' of mathematics and logic) — they function as representations, and those internal states as representational vehicles. As Haugeland (1981, p. 44) famously put it, "if you take care of the syntax, the semantics will take care of itself".

In sum, the central insight is that the notion of computation sheds light on how state-transitions in physical systems, though sensitive only to some of the physical properties of their internal states, can nonetheless respect semantic constraints. Computation gives us a way of understanding how to mechanise cognitive processes, including representational ones, in a way that is compatible with scientific materialism. To once again borrow effective terminology from Haugeland (1981, 1985), the notion of computation helps us to see how syntactic engines, *i.e.* systems governed by rules sensitive only to physical properties, can also be semantic engines, *i.e.* systems whose internal states represent, and whose state-transitions respect semantic constraints.

We see therefore how come the notion of computation comes to compose, with that of representation, the pillars of the cognitive sciences. Similar puzzles, however, arise for computation as they do for representation. We need to give an account of computation that is able to explain how computations take place in the physical world; and which is able suitably to distinguish computational processes from non-computational ones in such a way as to do justice to the practices of computer and cognitive scientists, as well as to the explanatory role of the notion in those sciences. We need, in other words, adequately to naturalise computation, that is, to explain in purely scientifically-acceptable terms what computations are, and how they are instantiated in concrete physical systems — giving a theory of computation in physical systems, or concrete computation. That will be the mission undertaken by Part II.

With a suitable account of concrete computation, cognition becomes amenable to causal explanation. If cognitive systems, inclusive of their representational properties,

[3]This, as many of the introductory remarks in this section, have been disputed in the literature. See, for a treatment of this issue, Rescorla (2012*a*).

can be understood as computational systems; and if computations and computational systems can find, courtesy of a theory of concrete computation, their place in the scientific worldview; we can explain cognition by means of computational operations over internal states, some of which representational vehicles, in a scientifically-acceptable fashion. This is the bet underlying most of the cognitive sciences: that naturalistic notions of computation and representation will provide the grounds to a materialistic, scientific, causal explanation of how cognitive systems work, casting away the mystery that surrounds cognition.

1.3 Some useful distinctions

The distinctions between representational and intentional content, and between cognitive and intentional states introduced above are not the only ones that I invite the reader to keep in mind throughout this work. A couple of other distinctions will frequently be at centre stage in my discussion; to keep any future confusion or puzzlement safely at bay, it is best to spend some words to clarify them from the get-go.

A contrast that will play an important role in what is to come is that between views that insist that representation, content, and computation are objective, observer- or mind-independent notions; and views that in contrast have it that some of those notions, or all of them, play a mostly pragmatic role, being dependent on our explanatory purposes, as well as on their heuristic value — *i.e.* their capacity to allow us better to understand events and processes in the world due to their capacity to simplify, make salient, or more synthetically and intuitively to describe phenomena. I will refer to the former — somewhat idiosyncratically — as objectivist views; whilst I will call the latter pragmatist views[4]. One may be an objectivist about one or more of those notions, and a pragmatist about the others. A position that I will put forward in section §8.1 is objectivist about concrete computation, while pragmatist about cognitive representation and representational content[5].

For an objectivist about representations, contents, and/or concrete computations, such notions earn their explanatory keep because they are fruitful scientific posits, and their positing is justified by their capturing some observer- or mind-independent feature of the world — *i.e.* independent of their being interpreted or understood by sentient beings, as well as the latter's explanatory purposes and practices[6]. Even though those notions are normally posited in the course of attempts to explain cognitive phenomena, their explanatory value is supposed to hinge on their corresponding, to some degree of accuracy, to features that cognitive systems possess, and would possess even in the

[4]Keep in mind that the foregoing distinction is not equivalent to the realism/anti-realism dispute. I will try and keep my treatment at a safe distance from the latter debate. Note moreover that my usage of the term 'pragmatist' does not refer to so-called neo-pragmatist theories of representation and content (*e.g.* Haugeland 1990, Cash 2009). Such views, which are not widely shared or discussed despite their considerable interest, will unfortunately also not be tackled in this work.

[5]Egan (2014*b*, 2015) also defends a view on these lines.

[6]Of course, those notions are not independent of sentient beings and their cognitive states insofar as they are meant to help explain exactly what cognitive states are, and how they work. But this is not the kind of mind-independence that is at issue here.

absence of any sentient being having cognitive states about those features, as well as the explanatory purposes and practices in which those notions are employed.

In contrast, for a pragmatist about representations, contents, and/or concrete computations, these notions depend on conceptual apparatus that we as cognitive scientists or users of folk psychology impose on the world, and which are not committed to their being faithful to its nature and organisation. This view of pragmatism is in line with proposals by Blackburn (2010) and Egan (2014*b*). As Blackburn (2010, p. 2) puts it, pragmatism about a notion or discourse involves three tenets:

1. you offer an explanation of what we are up to in going in for this discourse,

2. the explanation eschews any use of the referring expressions of the discourse ... or any semantic or ontological attempt to 'interpret' the discourse in a domain...

3. the explanation proceeds by talking in different terms of what is done by so talking ... [giving a story] about how this mode of talking and thinking and practising came about, and the functions it serves.

In the kind of pragmatism we are here concerned with, it is the functions a certain notion serves in our explanatory practices that are of particular relevance. Representations, contents, computations may be ascribed to systems as part of our explanatory practices in order to simplify explanations, to allow us more easily to grasp interesting connexions, or merely because regarding some systems as representing or performing computations proves to be heuristically useful in helping us make sense of the goings-on in cognitive systems. In such pragmatist views, the positing of representations, contents, computations is done with an eye to the fruitfulness and heuristic value for us to use those notions, rather than to capturing faithfully some observer-independent feature of the world. According to pragmatists, in other words, it may well be that the notions of representation, content, or computation do not capture anything that cognitive systems possess independently of the explanatory interests and practices that brings us to use them in our everyday or scientific understanding of the world. To echo once again Blackburn (2010, p. 12), when we employ such notions "we can see ourselves as having enriched our inferential practices of our dealings with the world, without having licensed the philosopher to enrich our conception of the world with which we are dealing"[7].

The foregoing distinction between objective and pragmatic views is though not nuanced enough to do justice to the rich debate on the nature and role of representation and content. A further level of complexity is introduced once we bring to bear considerations on their explanatory status in the cognitive sciences. Three traditional positions on this regard are often held to exhaust the conceptual landscape: robust reductionism, primitivism, and eliminativism. The two former are safely on the objectivist side of the divide, while the latter admits a double reading.

[7]Pragmatism thus characterised does not exhaust the alternatives to objectivism, as the following discussion will show. One interesting non-objectivist approach that I will not examine in this work is fictionalism, carefully explored by Sprevak (2013).

Roughly, robust reductionists hold that representations and representational contents are objective features of cognitive systems, which can be given reductive explanations in purely naturalistic and non-representational terms (*e.g.* Fodor 1975, Dretske 1981, Millikan 1984). On this view, some inner states of cognitive systems have representational content due to their standing in some special natural relation to what they represent. Those special natural relations bestow fairly determinate contents on cognitive states and processes; and representations are taken as structures in the cognitive system whose boundaries are relatively clearly definable, are stable, repeatable (*i.e.* participate in the same way in different cognitive processes, and have the same content), and often composable (*i.e.* that can be systematically combined with other representations to yield more complex representations).

Primitivists, in their turn, hold that the notions of representation and content play a crucial role in the successful and fruitful cognitive sciences, and this suffices for taking them as objective features of cognitive systems (*e.g.* Burge 2010). They are scientific primitives, which need not be themselves further explained, or naturalised. Primitivists take representation to be a scientific posit in no need of naturalisation, as much as fundamental entities posited in physics do not call for naturalisation, insofar as they are part of a successful and fruitful science.

Eliminativists, finally, believe that our best cognitive sciences will do away with appeal to representation and content, producing instead explanations at a purely functional or neurophysiological level. Eliminativists hold that propositional attitudes and intentional content are probable candidates for elimination from future cognitive science (*e.g.* Churchland 1981, Stich 1983), while more radical eliminativists have claimed that even representational content will be eliminated (*e.g.* Van Gelder 1995, Hutto & Myin 2013). The question of how to give a naturalistic reduction of representation thereby does not exercise the eliminativist, though they must show that equally satisfying explanations of cognitive phenomena can be provided without appeal to representations.

Eliminativism can be read as an attack on the objectivist view of representation and content, or on the pragmatist view, or on both, depending on how strongly the eliminativist claim is interpreted. The claim may be that the notions of representation and intentional (or, more radically, also representational) content do not capture objective features possessed by cognitive systems. This claim is compatible with a pragmatist view, for we may want to keep talking of representations and contents, at least in some domains of our worldly practices, due to their heuristic value in helping us to comprehend the workings of cognitive systems or of fellow sentient beings[8]. But the eliminativist, I believe, should be read as making a stronger claim: the notions of representation and (intentional) content not only do not capture anything objective about cognitive systems, but they are not even explanatorily useful in the cognitive sciences. It is due to their putative explanatory uselessness — or even worse, their power of misleading research — that they should be eliminated from the cognitive sciences. That is

[8] Among moderate eliminativists, the Churchlands seem to endorse a version of such a view. Propositional attitudes, by their lights, can still be usefully ascribed to people and perhaps animals in our everyday, non-scientific dealings with the world.

to say, the eliminativist denies the pragmatist view as well.

In contrast to mainstream lore, it will be the burden of this work to show that there are other stable, interesting, and even attractive positions in addition to the three presented above. In chapter 8, I will show that there are at least two further plausible views on the nature and explanatory role of representation and content in the cognitive sciences, one on each side of the objectivist/pragmatist divide.

From one side, one may argue that representation and content should be seen as ineliminable parts of cognitive science, *contra* the eliminativist; but nonetheless reject the claim that those notions capture anything objective in cognitive systems — *contra* reductionists and primitivists — content being rather part of an explanatory gloss dependent on pragmatic considerations. This view, content pragmatism, has been held most prominently by Egan (1999, 2009, 2010, 2014*b*). It is pragmatist about representation and content, but not eliminativist.

On the objectivist side, one may agree with robust reductionists and primitivists that representation and content are objective features of cognitive systems, but deny that representational vehicles have the properties that robust reductionists believe they do, such as stability, repeatability, compositionality, relatively clear boundaries, and fairly determinate content; as well as deny that there is a relatively bounded set of natural relations that determine content. Moreover, one can reject both the primitivist claim that no naturalisation of the notions of representation and content is required, and the terms set by the robust reductionist as to how such an account will have to look like. Such a mild reductionist view, inspired by Dennett (1987*a*, 1991), has it that representation and content capture fundamentally variable, context-sensitive, and disjunctive features of cognitive systems, for which no straightforward naturalistic reduction is possible. In other words, one may be an objectivist about representation and content, though a non-robustly reductionist one.

In summary, views about computation, representation, and content can be distinguished in light of whether they take those notions to capture objective, observer-independent features of the world (objectivism); or, on the contrary, regard them as notions that depend fundamentally on our explanatory practices and aims, on what we find useful in making sense of the world given our cognitive endowment and interests, regardless of whether they capture observer-independent features (pragmatism). Theories of representation and content, in their turn, can be further distinguished in terms of the explanatory status they bestow on those notions in the cognitive sciences. I have briefly presented three mainstream views on this regard: robust reductionism, primitivism, and eliminativism. Though they are often taken to exhaust the conceptual possibilities, I will argue that this is not so. Alternative views, such as content pragmatism and mild reductionism, are cogent and worth of investigation.

Treatment of these alternative positions will have to wait for chapter 8. However, I invite the reader to bear in mind that pretty much everything that follows is meant to work as grounds and motivation for exploring and helping develop such alternative views. Let us then begin the slow build up toward Part III. We shall start with a

brief overview of the robust reductionist project, which has had pride of place in the literature on representation for the past four decades or so.

1.4 Naturalising content: the project

Robust reductionism has plausibly been for the past decades the mainstream view of the nature of representation. Philosophers, especially from the 1980's on, have put forward, examined, criticised, and tweaked theories of representation and content that attempt to ground these notions on a constrained set of purely naturalistic relations between representational vehicles, and entities and processes external to the cognitive system. This project, which spun a vast and rich literature, came to be known as 'naturalising intentionality'. Given the terminology I am employing in this work, I will prefer the expression 'naturalising content'. Theorists interested in this naturalistic endeavour often did not limit themselves to the contents of intentional states, but also tackled issues relating to the representational content of cognitive states — though frequently treatment of the latter was seen as little more than a means to get to the more complicated case of intentional states, and even the foregoing distinction was only occasionally acknowledged.

The naturalising content project has been only cursorily interested in issues relating to the format of representations, and what makes them function as representations, though it has had effects on those debates. The focus of the endeavour has been that of explaining, exclusively by means of scientifically-acceptable, naturalistic, objective relations, how representations get their contents, *i.e.* how physical states of cognitive systems come to bear semantic properties[9]. The guiding idea is to reduce content, or semantics, or aboutness, to naturalistic factors and relations. Paraphrasing Fodor's (1987) apt phrase, if there is content, it must be really something else. Such reduction must be non-circular. As Fodor (1984, p. 232) puts it, what the robust reductionist wants to get with their naturalising project is "... at a minimum ... something of the form 'R represents S' is true iff C where the vocabulary in which condition C is couched contains neither intentional nor semantical expressions".

A satisfying naturalistic theory of content must respect further constraints: it must provide an account of representation that is relevant for the cognitive sciences; and which is able to play the required explanatory role. Particularly important consequences of these further requirements is that the contents bestowed on representations must be fairly determinate — excessively vague or disjunctive contents jeopardise the explanatory role of representation — and that the theory must make space for the possibility of misrepresentation, *i.e.* cases in which there is representational error. The latter requirement is normally taken to be essential for a satisfying theory of content[10]. As we have seen in section §1.1, what defines content is some form of correctness conditions, broadly understood. This very characterisation tinges the notion with a kind of

[9] See Ryder (2009*b*) for a thorough review.

[10] Though the essentiality claim has been denied by some. See Cummins (1996), Perlman (2000), Isaac (2012).

normativity. There is something it is for a representation to be true or false, satisfied or else, accurate or inaccurate, etc. Thereby theories of content must make space for the possibility of misrepresentation, on pain of failing to be theories of content.

The possibility of misrepresentation is germane for the explanatory purchase of the notion of representation in explaining behaviour. As much as we want to explain complex successful behaviour by means of correct representation, we also want to explain failures in behaviour by means of incorrect representation. The problem of misrepresentation is closely connected with the quest to provide a theory that bestows fairly determinate contents on representational vehicles. For if contents are too indeterminate or disjunctive, it is mysterious how they can be wrong — or right. Unsurprisingly, a large part of the efforts carried out by philosophers invested in the naturalising content project has been dedicated to figuring out how to account for misrepresentation. This is far from a trivial puzzle: for such an account to be had, we need to extract the required sort of normativity out of the raw materials that the natural, objective world makes available.

A further *desideratum* sometimes proposed is that the notions of representation and content that fall out from naturalistic theories not be too liberal — that is to say, that they do not make too many things be representations. Most accounts on offer would run afoul of excessive liberality: too many states, cognitive and non-cognitive, end up counting as representations (and thus as being contentful), putatively jeopardising the distinctive explanatory role of the concept in the cognitive sciences[11].

In this section, I will briefly present the four main candidate naturalistic theories of content, all of which have seen the light of day in the 1980s, and all of which belong to the robust reductionist camp: functional role semantics; two influential types of informational semantics, namely causal-informational semantics and teleoinformational semantics; and teleosemantics. The next section will present some of the traditional objections, especially for what regards the problem of indeterminacy of content, that have been moved against these views.

Please bear in mind that my treatment of this complex debate will be rough and sketchy: my aim is just to provide an overview of the theoretical background against which my positive proposal, developed in Parts II and III, is to be developed. The next chapter, in its turn, will be concerned with a theory that has been a sort of underdog in this debate: structural representation.

Before we move on, one note: though the naturalising content project is normally associated with robust reductionist views of representation, it is not the case that only robust reductionism is concerned with providing naturalistic conditions for determining the contents of representations. As we will see in Part III, a mild reductionist view also aims at providing such conditions, though in a way importantly different from robust reductionism.

[11] See Ramsey (2007), Burge (2010), Morgan (2014).

1.4.1 Functional Role Semantics

Functional role semantics, also known as inferential and conceptual role semantics, has had some currency in the 1980s (see, *e.g.*, Harman 1982, Block 1986), and though it has somewhat fallen out of favour as a general theory of representation and content since, it still has proponents (Churchland 1998, 2012). According to functional role semantics, the causal role that a certain state plays in the economy of the cognitive system or beyond is what determines its content. Causal role is normally understood as inferential or conceptual — that is to say, it is the role a state has in licensing, and being the outcome of, inferences from and to other states in the system. Cognitive states appropriately cause or are caused by the tokening of other states to which they are inferentially connected.

Functional role semantics has internalist and externalist versions. Internalist, or short-armed views, have it that only the functional roles internal to the cognitive system count as content-determining, while externalist, or long-armed views, include causal connexions to the environment. Internalist views of functional role semantics are normally meant to be part of a more complex theory of content that also includes an external, often causal-informational, factor — leading to two-factor theories of content (Block 1986). Theories of this type tend to share the advantages, as well as the problems, that come with the appeal to each of the factors.

1.4.2 Informational Semantics

Informational semantics, put forward by Dretske (1981) and further developed in different ways by many others (*e.g.* Fodor 1987, Dretske 1988), bases representation on indication, or information-transmission. Roughly, the idea is that a state r is a candidate for being a representation if it carries information about some other state t. According to Dretske, a state r carries information about another state t if its having a certain property F raises to 1 the probability of t having a certain property G, given certain background conditions, which Dretske calls 'channel conditions'. Put in other terms, r being F is a natural sign for t being G, as a tree stump having a certain number of rings signals — or carries information about — how old the tree is. There is a correlation between the two states, the former covaries with the latter. In Dretske's terminology, the former state indicates the latter.

The limitation to natural signs is meant to rule out signs that depend on interpretation or intention, such as words and traffic signs. These nonnatural signs, or symbols, carry information about states of affairs only by passing through interpretive systems, and do not depend, in opposition to the tree rings, simply on a natural relation between sign and signified. Informational semantics purports to explain representation by recourse to natural signs and signifieds. The assumption, which is at the basis of the naturalising content project, is that "there is something *in* nature (not merely in the minds that struggle to comprehend nature), some objective, observer-independent fact or set of facts, that forms the basis of one thing's meaning or indicating something

about another"[12].

The natural relations that ground indication are normally causal relations[13]. Signs normally indicate — or carry information about — what caused them. For this reason, informational semantics is regarded as a type of causal theory of content. However, there need be no direct causal connexion between indicator and what it indicates, for instance if both are generated by a common cause. As a theory of representation, informational semantics needs to account for how representational content gets fixed, and for how misrepresentation is possible. Indicators carry information about many different states of affairs, thus having non-unique informational contents; while a theory of representational and intentional content aims at fairly determinate contents. Indication moreover does not leave space for misrepresentation, for there is no misindication — there is no sense in which a state can incorrectly carry information about something[14]. More is needed if informational semantics is to become a full-fledged theory of representational content. Dretske has formulated two variations of the view, one purely based on causal-informational relations (Dretske 1981), another including teleological factors (Dretske 1986, 1988). Let us take a brief look at each in turn.

Causal-informational Semantics

In his early work on informational semantics, Dretske (1981) tried to give a complete theory of representation that relied exclusively on causal-informational relations between indicators and what they indicate[15]. In order to do justice to the determinacy of representational content vis-à-vis the multiplicity of informational content that each indicator bears, Dretske introduced the notion of digitalisation. There is no need (nor space) to look at the details, but the basic idea is that during learning the cognitive system selects, from the multiple informational contents carried by a certain indicator state, one piece of information (the most specific one), and that piece of information is the representational (or intentional) content of the state. In this way, the non-uniqueness of informational content can be arguably curbed in its transition to representationhood, partly effected by means of the process of digitalisation.

The most specific piece of information carried by an indicator state can never be false, given the definition of what it is for a state to carry information. To account for the possibility of misrepresentation further factors must be brought in. For Dretske (1981), the additional factor supposed to play the larger role in making space for misrepresentation is a circumscribed learning period, during which the content of cognitive (and intentional) states gets fixed. During the learning period, the conditions for cor-

[12] Dretske (1988, p. 58.)

[13] Dretske (1988, p. 56.)

[14] See Dretske (1988, p. 56), Fodor (1984, p. 239.)

[15] Another purely causal-informational theory was developed by Fodor (1987, 2008), who tries to provide a logical condition to distinguish correct from incorrect tokenings of a representation, thus making space for misrepresentation, while remaining faithful to a purely causal-informational view — the asymmetric dependency theory. Fodor's theory is purely formal: rather than a full picture of how there comes to be misrepresentation, it is a formal requirement on any theory of error — it remains silent on the mechanisms behind the relevant dependencies. The proposal has been subjected to several objections (Cummins 1989, 1996, Prinz 2000, Dunlop 2004, Rupert 2008).

rect indication are optimal: channel conditions are appropriate, there is the guidance of some sort of teacher, feedback is adequate, etc. When the learning period ends, representational (or intentional) content gets fixed once and for all. All subsequent tokenings of the now fully representational state not generated by its content count as misrepresentations. Misrepresentation takes place when a representation, whose content was fixed during the learning period, happens subsequently to be tokened when there are no causal-informational relations in place with its content.

Teleoinformational Semantics

In his more mature version of informational semantics, Dretske (1986, 1988) turned to teleology[16]. Instead of a learning period that fixes representational content, Dretske has recourse to functions indicators have been recruited for during the individual development of the cognitive system. During development, a kind of associative learning takes place which selects indicators that carry information about some state of affairs O in the world and endows them with the function of providing information about O. This selection, or recruitment, depends on the successful use of the indicator in guiding behaviour towards O.

Behavioural responses are associated with indicators that carry information about the stimulus condition in which that kind of response brings forth rewarding outcomes. By means of associative reinforcement these indicators are gradually recruited by the system as causes of appropriate behaviour. The indicator becomes a representation of the state of affairs O in which the use of that indicator in guiding behaviour leads to success — the indicator is recruited as a representation of those conditions that make the behaviour appropriate — while all the other states of affairs that it indicates are excluded from its representational content. Misrepresentations are those occurrences in which representations are tokened by something that is not what they have ontogenetically acquired the function to represent.

Basing representational content on a notion of teleological function — *i.e.* on the purpose, or end, of systems — has the advantage of providing a naturalistic means of accounting for the normativity of content. Systems that have functions can go wrong: they can fail to perform their functions. Teleological theories of content generally rely on evolutionary history and/or learning to account for how there can be such things as teleological functions in the world[17]. Theories of content that rely on teleology are perhaps the most promising robust reductionist approaches to representation and content on offer. In Dretske's version of teleoinformational semantics, learning plays a central role. Other theories give pride of place to natural selection, as is the case of Millikan's version of teleosemantics, to which we now turn.

[16] Other influential teleoinformational theories have more recently been put forward by Shea (2007), Neander (2013).

[17] I survey theories of teleological function in section §5.3.

1.4.3 Teleosemantics

Teleosemantics has been most forcefully put forward by Millikan (1984, 1989*a*, 2004). The guiding idea is to rely on teleology, and in particular teleological functions arisen from evolutionary processes, to account for representation and content — abandoning the Dretskean idea that information-carrying is at the basis of representationhood. As Neander (2013, p. 22) points out, while informational semantics focuses on the causes of representations (or what they carry information about), teleosemantics focuses on the effects that representations have. The burden shifts from what produces a representation to how it is used (or differently put, from the representation producer to the representation consumer — whereby the label 'consumer-based teleosemantics' applied to Millikan's view).

According to teleosemantics, what determines the content of a representation is the use that is made of it by a consumer system. That use explains the persistence of the consumer system in the phylogeny, *i.e.* its having been selected by evolution for doing the specific thing it does[18]. This history of selection is what bestows teleological function on such systems. It is because a representation was used often enough in a certain way — leading often enough to behaviour that was adaptive — that the consumer system that used it in that way got selected, and thereby persisted in organisms across generations. The content of a representation consists in the conditions that led the consumer system to function adaptively by means of using the representation. Put differently, the contents of a representation are the conditions that make it so that the way it affects the consumer allows the consumer to perform its teleological function, *i.e.* that leads to behaviour that is adaptive, and that specifically explains the persistence of the consumer system in the phylogeny.

There are several variations on the basic insight of teleosemantics, giving rise to different theories of content which share the reliance on teleological function as playing a central role in the fixation of representational content. The quite modest aims of this overview do not call for a closer examination of such theories.

1.5 Naturalising content: the obstacles

The mainstream robust reductionist theories of content that we have briefly examined above have attracted as many detractors as they have adherents. Critics have pointed out several shortcomings of the proposed naturalistic reductions of representational (and intentional) content. Many of the objections moved belong to two distinct lines of criticism.

One line accuses the mainstream theories of representational content of being too liberal: they ascribe representational content to too many things in the world, failing therefore to account for, or at least considerably watering down, the peculiarly psychological interest of the notion[19]. However, when it comes to subpersonal states and

[18] Somewhat analogous formulations can be given for ontogenetic learning-based selection.

[19] See, for this kind of criticism, Burge (2010), Morgan (2014), as well as chapter 8.

representational content, it is doubtful that this line of objection has much traction. Even if representational content may be ascribed to non-cognitive states and processes as well as to more peculiarly cognitive ones, further argument must be offered to show that this jeopardises the explanatory role of the notion in the cognitive sciences. At least when it comes to subpersonal states and their representational content, it is not obvious that there is a principled distinction between the nature of the processes and interactions at play in cognitive systems, and outside them. On the contrary, I take it to be plausible that there will not be a real distinction here. It is likely that there will be one such distinction when it comes to intentional states and intentional content, phenomena arguably proprietary to psychology. Theories of representational content need not thereby worry about casting the representational net widely, capturing cognitive and non-cognitive systems alike.

A second type of difficulty for robust reductionist views of content is more relevant for our purposes, and has largely dominated the literature on mental representation: the naturalistic reductions they offer may fail to yield determinate contents, making misrepresentation problematic, and jeopardising the explanatory role of the notion in the cognitive sciences. Indeterminacy objections are supposed to show that theories of content end up bestowing disjunctive contents on representations. Instead of determinate, unique or fairly unique contents, representations would have as their contents a disjunction of sundry states-of-affairs — instead of representing X, as we would want, representations would end up representing X or Y or Z or ... having thereby non-unique, or indeterminate, contents.

Indeterminacy of content objections are sometimes taken to exhaust what Fodor (1984) has called 'the disjunction problem'. The disjunction problem arguably shows that causal-informational theories of content make no space for misrepresentation at all, yielding indeterminately disjunctive contents. Indeterminacy problems need not all be so severe: theories of content may yield limitedly disjunctive contents, introducing considerable vagueness as to which cases count as instances of misrepresentation. In this sense, the disjunction problem is a particularly severe kind of indeterminacy problem. I will stick to the expression 'indeterminacy problems' to cover all objections from indeterminacy of content, regardless of their degree of severity.

In this section, I will briefly review traditional arguments that have been moved against the mainstream theories of content presented above. These arguments have spawned a rich and variegated debate, which led to considerable refinements of the theories of content they attack, as well as of the attacks themselves. I cannot here but scratch the surface of such debate. For my purposes, at any rate, a detailed treatment is not needed. My aim is just to point to issues and questions that theories of content, including the ones I put forward in Part III, have to tackle, either by providing convincing answers to them, or by keeping them from arising.

1.5.1 Holism

In functional role semantics the causal-inferential role played by internal states in the economy of the cognitive system, *i.e.* their abstract 'position' in a network of internal states, helps determine their contents. Content-fixation is thereby dependent on the whole set of cognitive states and their relations, making functional role semantics a holistic theory. Holism is problematic for a theory of representation and content, for it makes sameness of content virtually impossible across different cognitive systems (or across the same cognitive system in different times, provided there was learning in between)[20]. If contents are determined by the whole network of cognitive states, and given that any two cognitive systems are extremely likely to be different in the number and functional role of at least some of their states, it follows that two cognitive systems (or two time-slices of the same system) will never share contents — which, beside being implausible, jeopardises the possibility of communication and the stability of content.

One line of reply to the problem of holism is to insist that only parts of the network of internal states fix the content of a given cognitive state (molecularism). However, it is difficult to give a principled means to distinguish between the content-determining and the non-content-determining parts of the network of causal-inferential relations, in particular without helping oneself to the controversial analytic-synthetic distinction. Another line of reply has been to concede that any two cognitive systems will never share fine-grained content, but that for all purposes what matters is content similarity, rather than identity (Churchland 1998). It is controversial whether there are principled ways to measure content similarity, and how much similarity is needed to allow sharing of intentional states [21].

1.5.2 Causal explanation

One *desideratum* for theories of representation and content is that they make it so that representation and content can play a role in the causal explanation of successful and unsuccessful behaviour. The determination of content and representational status should ground the causal relevance of representations to the explanation of behaviour. In other words, representations lead to the behaviours they do due in part to their contents, rather than having the contents they do due to the causal consequences they lead to. Functional role semantics seems ill-equipped to deal with this *desideratum*. By its lights, it is the causal powers of vehicles that help bestow content on them, making causal explanation of behaviour in terms of content circular.

1.5.3 The problem of error

The normativity of representational content requires that theories of content make space for the possibility of representational error, or misrepresentation. Fodor (1984) argues that causal-informational theories, such as Dretske's (1981), are ill-equipped to respect

[20] See Fodor & Lepore (1992).

[21] For criticism of such a strategy, see Fodor & Lepore (1999).

that requirement. Suppose that a representation R carries digitalised information, and thereby represents X, say penguins. Suppose further that occasionally instances of Y, say puffins in a dark night, also cause tokenings of R. Tokenings of R generated by puffins-in-a-dark-night are what Fodor calls 'wild tokenings'. Wild tokenings are the primary cases that causal-informational theories would want to rule as instances of misrepresentation — a representation that has as its content X, by some error or failure, perhaps due to less than optimal channel conditions, comes to be caused by something else, therefore misrepresenting how things are in the world (*e.g.* representing a puffin as a penguin).

However, causal-informational relations by themselves cannot make space for misrepresentation: there is no misinformation, as we have seen, and there is no (non-intentionally-laden) sense in which a cause generates its effect by mistake. If puffins-in-a-dark-night can cause tokenings of R, it follows that they stand in appropriate causal-informational relations to R as much as penguins, and are therefore to be included in its content — R would in consequence have as its content penguin or puffin-in-a-dark-night or Z, where Z stands for the disjunction of all other states-of-affairs that can cause R. Causal-informational theories, unassisted, cannot thereby make space for misrepresentation, for everything that may cause a tokening of the representation will have to be included in its content, leading to wildly disjunctive contents — all tokenings, even wild ones, will count as cases of correct representation[22]. No misrepresentation is hence possible.

Dretske's (1981) attempt to circumvent this issue is unsatisfying. He appeals to a learning period, during which channel conditions are optimal (thus putatively avoiding wild tokenings), and content is fixed once and for all. After the learning period is over, wild tokenings would be cases of misrepresentation: what causes the tokening of the representation is not its content, *i.e.* the type of cause that tokened the representation during the learning period.

Fodor (1984) exposes three fatal flaws with the foregoing. First, there is no principled way of fixing the boundaries of the learning period — cognitive systems are in constant interaction with their environment, and learning takes place during the whole lifetime[23]. Second, even granting that the notion of a circumscribed learning period is sensible, the account excludes, or is at least silent on, representations that are innate, that is to say, that are not learned. Third, the solution does not take into account relevant counterfactuals. Since Y, a wild cause, is sufficient to token R, were it to have been present during the learning period, it would have caused R. In that case, the content of R would have been disjunctive, *i.e.* X or Y or Z. Thus, when the relevant counterfactuals are taken into account, the appeal to a learning period fails to solve the problem of error: every wild tokening still counts as a case of correct representation, and misrepresentation

[22] A similar worry applies to long-armed versions of functional role semantics.

[23] A related worry regards the assumption that channel conditions are optimal during the learning period. It seems impossible to specify what the optimal channel conditions are without already presupposing the contents of representations. For instance, the channel conditions that are optimal for visual perception are not the ones that are optimal for auditory perception.

is impossible[24].

Causal-informational theories need therefore to appeal to other supplementing factors in order to curb indeterminacy of content, and make space for misrepresentation[25]. The problem of error is the first indeterminacy problem presented here, the other three follow in the next three subsections.

1.5.4 The distality problem

An issue that arises for all the popular theories of content presented above is the distality problem (*cf.* Dretske 1986). Normally there are many processes and states that establish the connexion between cognitive systems and their environments, some of which are content-fixing. But most of those intermediate states and processes, it is held, should not be part of the content of representational states. A representation of a visually perceived distal object, for instance, should not have among its contents the patterns of light impinging on the retina that are part of the causal chain leading from the distal object to the cognitive system. Theories of content should have the means to exclude these intermediate states and processes from appearing in the contents of representations of distal objects. Otherwise, representational content becomes indeterminate: representations would represent the distal object as well as all the intermediate states and processes that connect the distal object to the cognitive system — including lower-level representational states inside the system.

Solutions to this problem have been proposed (*e.g.* Dretske 1986), but it is unclear whether and to what extent they succeed. It is often held that teleosemantics is able to solve the distality problem, though this has been put into doubt on the grounds that consumer systems have been selected in part because they respond adaptively to the intermediate states and processes that link the distal object to the cognitive system, since these perforce mediate the relationship between world and cognitive states (Neander 2012).

1.5.5 Functional indeterminacy

One of the most common and most widely discussed problems with theories of content that rely on teleological considerations is functional indeterminacy. The worry is that teleological functions have themselves a degree of indeterminacy which infects the representational contents they are supposed to help bestow on cognitive states. A related observation points out that teleological functions are not fine-grained enough to allow the bestowal of determinate contents. From these types of functional indeterminacy it would follow that theories of content that rely on teleology end up with indeterminate representational contents.

[24]This third objection only applies to causal-informational theories, such as Dretske's, that appeal to dispositions, rather than actual history. If it is the dispositions to token representations during the learning period that are relevant, rather than actual tokenings, counterfactual considerations become relevant. A causal-informational theory that appeals instead to actual history has been put forward by Prinz (2002). See Artiga (2014) for critical discussion.

[25]More recent, probabilistic causal-informational theories hold some promise on this regard (*cf.* Usher 2001, Eliasmith 2005).

Some authors (*e.g.* Cummins 1989, Fodor 1990) have pressed the point that there is indeterminacy between conservative and liberal interpretations of the teleological functions responsible for determining content in teleoinformational and teleological theories. Famously, the question 'does the representation that brings the frog to snap its tongue and successfully capture a nutritious fly represent "fly", "food", or "moving dark spot"?' may not have a straightforward answer.

Dedicated neural circuitry in frogs generates snapping of their tongues when it detects small moving dark spots in their visual fields. Given that in frogs' usual habitat small moving dark spots happen to be edible insects, this circuitry has been plausibly selected by evolution given its adaptive value. All three contents mentioned above, 'fly', 'food', and 'moving dark spot' would have been appropriate to guide successful behaviour, given the environments where frogs live and thrive. It is not clear, in consequence, which the determinate conditions are that allow consumer systems in the frog, such as its stomach, to fulfil their teleological functions: is it presence of flies, food, or moving dark spots (or something else yet)? And, pushing further, what are the teleological functions of consumer systems, such as the frog's stomach? Is it to digest flies, food, or nutritious moving dark spots?[26] Teleology arguably lacks the means of picking one of these conditions as bestowing representational content in exclusion of the others, yielding thereby indeterminate representational contents.

1.5.6 Quinean indeterminacy

The final of the four indeterminacy problems I will present here stems from the famous problem of indeterminacy of translation formulated by Willard van Orman Quine. Though Quine's point was aimed at problems surrounding linguistic translation and interpretation, similar issues arise for theories of representational content. Briefly, theories of content do not seem equipped to distinguish different, but co-extensional contents. For instance, how to determine whether a certain representation represents 'rabbit', 'rabbit time-slice', or 'undetached rabbit parts'? Since such contents are co-extensional in all possible cases, there are seemingly no natural relations that can allow theories of content to fix one of the co-extensional contents as the determinate content of a representation (Gates 1996). As a consequence, such theories can at best bestow disjunctive co-extensional contents to cognitive representations.

Quinean indeterminacy is a very general problem, which may undermine the notion of causal explanation itself by making the causal *relata* of any causal relation indeterminate between co-extensional properties. Given the fact that it is not specific to theories of content — in contrast to the indeterminacy problems presented above — Quinean indeterminacy is of secondary importance in this debate.

1.5.7 Cummins and Burge against teleosemantics

Cummins (1996), and more recently Burge (2010) purport to undermine the idea that

[26] An analogous problem applies to producer-based teleosemantics: the teleological functions of representation producers are indeterminate between (at least) these three conditions.

teleology, be it dependent on individual learning or natural selection, can be relevant for fixing representational content. They argue that the success on which biological notions of function hinge is behavioural success, and not semantic, or representational success. The success conditions that lead to 'recruitment' or adaptive success are not conditions on correct representation. Inaccurate representations may lead to behavioural success, and they may even be better in assuring such success than accurate representations. The latter may often be too cognitively expensive and slow to generate, hindering behavioural success rather than promoting it.

Behavioural success and representational success, even though they may coincide in some (or even most) cases, cannot be equated. Hence we cannot use behavioural success to ground representational success. Acquired functions of representational mechanisms, based as they are on behavioural success, cannot provide conditions on correct representation. The mechanisms may fulfil their function and lead to behavioural success, but this does not say anything about whether the representations tokened are correct or else: they may be wrong, but 'good enough'. As Burge (2010, p. 302, n. 18) puts it, "one cannot assimilate issues of accuracy and inaccuracy to issues of practical use ... functioning to be accurate is not *in itself* a biological function, at any level ... Biological functioning is not a semantical matter ... It is a practical matter, a matter of fitness for procreation".

Teleological considerations — so the argument goes — cannot be used, *contra* teleosemantics, to ground a theory of representational content. Teleology is not suitable for the task because what generates appropriate behaviour need not represent the world correctly[27]. Teleological approaches to content would run together two types of correctness, behavioural and representational, that are independent of one another, and though they may coincide in some cases, they need not so do.

1.6 Representation and the challenge of indeterminacy: new paths

Theories of content, if they are to be satisfactory, must have the tools appropriately to deal with the worries briefly examined above. Mainstream theories have gone through considerable refinements in order to try and meet those objections. It is not my aim to assess whether these improvements adequately address the worries. Rather than examining the prospects of existing theories, my objective in the foregoing is putting forward a new, substantially different account of representational content. A theory in which most of the above problems do not arise, or are easily and without much ado dealt with, requiring the addition of no epicycles. A theory, importantly, that is able to yield a notion of representation and content able to play the required explanatory role in the cognitive sciences.

My claim will be that a deflationary theory of content is just such a theory. The search for more robust theories of representation may turn out to be unnecessary for the

[27] See also Cummins (1996, p. 57).

project of securing the foundations of the sciences of the mind, avoiding metaphysical quagmires that can bog down more metaphysically-laden theories. I will take pains to show in coming chapters, and especially in Part III, that such a deflationary, and explanatorily adequate theory of representation can be had.

Before turning to the positive side of this work, it is worthwhile to examine an old robust reductionist theory of content that has been somewhat neglected in recent philosophical work. Despite its many shortcomings, I will later on show that such theory, structural representation, paves the way for a deflationary view of representation and content that has much to recommend it. This will bring this brief overview of existing theories, as well as Part I, to an end.

Chapter 2

Structural Representation

An intuitive idea about what representations are and how they work has it that representations resemble what they represent. It is by resembling their contents that representations are able to convey information about the world to the cognitive system — representations would be something not very distant from a copy, more or less accurate, of what is represented. By inspecting the representation, the cognitive system, or subsystems thereof, would be able to inform its behaviour in light of external circumstances. By having an internal copy of the world, the cognitive system is able to use it as a stand-in, a surrogate, to the world itself.

Since Antiquity, this intuitive idea has furnished the insight motivating many philosophical attempts to make sense of how representations work. The idea that representations take on, in some sense, the form of what they represent can be found in the works of thinkers such as Plato, Aristotle, and Hume[1]. Resemblance theories of representation, despite their historical popularity, have fallen out of favour during the 20th century, leading to their almost complete abandonment. An important role in the movement away from resemblance theories was played by Nelson Goodman's attack on such theories[2]. Goodman (1976) pointed out that resemblance has logical properties that representation lacks, the former being thus inadequate to ground a theory of the latter. In fact, while resemblance is a reflexive, symmetrical, and transitive relation, representation does not generally feature these relational properties — a representation does not represent itself, what is represented does not represent the representation, and if A represents B, and B represents C, it does not follow that A represents C.

Despite the apparently fatal blow against resemblance theories inflicted by Goodman, work on sophisticated versions of such theories have proceeded alongside the mainstream debates, which saw informational and teleosemantic theories take centre stage (Swoyer 1991, Gallistel 1990, Cummins 1996, O'Brien & Opie 2004, Bartels 2006, Isaac 2012, Shea 2014). I will refer to a subset of these refined theories of content, which follow on the footsteps of traditional resemblance theories, as theories of structural representation.

[1] See Cummins (1989) for a brief, and rather impressionistic overview.
[2] See also Suárez (2003).

2.1 What is Structural Representation?

Structural representation is usually cashed out as a resemblance theory of content (Cummins 1996, O'Brien & Opie 2004)[3]. Resemblance theories of content claim that a representation R represents an entity/event/state-of-affairs[4] O if R resembles O in certain ways and to a certain extent. Different resemblance theories fill up the details differently. The basic idea behind the structural representation view, or as Cummins (1996, p. 93) calls it, the "Picture Theory of Representation", is that representations represent by virtue of sharing relational structure with what they represent.

A way in which two entities can share relational structure is by means of structural resemblance. At its most general, as O'Brien & Opie (2004, p. 15) put it, "one system *structurally resembles* another when the physical relations among the objects that comprise the first preserve some aspects of the relational organisation of the objects that comprise the second". This sharing of relational structure, which can be quite abstract, allows representations to work as models, as stand-ins for what they represent. Such representations thereby make possible surrogative processing: one can operate on the representation, on the model, in order to draw conclusions about the things modelled.

Structural representations can be cognitive or non-cognitive. Maps, scale models, computational models, and many other non-cognitive representations qualify as structural representations, and are characterised partly by their ability of allowing surrogative processing. Such representations need not be static; quite on the contrary, the more interesting cases are those of dynamical representations. Dynamical models are such that the changes that the model undergoes or can undergo mirror the changes that what is represented undergoes or can undergo. A model of our solar system is a case in point. By changing the position of the globe that stands for a planet in the model, the relative positions of the other globes, which stand for the other planets, change accordingly. The relations between the globes represent the relations between planets of the solar system across different conditions.

> Structural representation enables us to reason directly about a representation in order to draw conclusions about the things that it represents. By examining the behaviour of a scale model of an aircraft in a wind tunnel, we can draw conclusions about a newly designed wing's response to wind shear, rather than trying it out on a Boeing 747 over Denver. By using numbers to represent the lengths of physical objects, we can represent facts about the objects numerically, perform calculations of various sorts, then translate the results back into a conclusion about the original objects. In such cases we use one sort of thing as a surrogate in our thinking about another, and so I shall call this surrogative reasoning. (Swoyer 1991, p. 449)

The story for cognitive representations is analogous. By sharing structure with what

[3] Though it need not be, since there may be structural correspondences between entities that are not based on similarity relations.

[4] For reasons of economy, I will drop this tripartite characterisation and use the term 'entity' to cover all three cases.

they represent, cognitive representations can be used by the cognitive system as models of the world. Possibly, such models, as in some non-cognitive cases, are dynamical, rather than static: transformations over the representational vehicles of a structured collection of representations mirror transformations that take place in what is represented. Representational vehicles represent elements of the represented domain, and the relations between those vehicles represent the relations between elements in the represented domain (Shea 2014).

At least in the case of cognitive representations, the insistence on the sharing of *relational* structure is crucial. The appeal to relational structure is a more precise and plausible way of cashing out the idea that representations take on, in some way, the form of what is represented. Relational structures are somewhat abstract, and can be implemented in different physical *substrata*. The relational organisation of elements in a structure is a second-order property, as it abstracts away from the physical properties of the elements themselves. For this reason, structural resemblance is a type of second-order resemblance.

A more precise definition of second-order resemblance has been put forward by O'Brien & Opie (2004, p. 11): given sets of objects V, O, and sets of relations R_{V}, R_{O} defined on the members of V and O, respectively,

> there is a *second-order resemblance* between two systems $S_V = (V, R_{\mathrm{V}})$ and $S_{\mathrm{O}} = (O, R_{\mathrm{O}})$ if, for at least *some* objects in V and *some* relations in R_{V}, there is a one-to-one mapping from V to O and a one-to-one mapping from R_{V} to R_{O} such that when a relation in R_{V} holds of objects of V, the corresponding relation in R_{O} holds of the corresponding objects in O.

From this definition it follows that second-order resemblance is widespread — many systems, be they entities or parts of entities, stand in second-order resemblance relations to many other systems. The definition puts no constraints on how to individuate systems, and requires only that *some* objects and relations in a system map onto objects and relations in the other. How these are chosen is left open. Moreover, there are no constraints on which kinds of objects and relations are at issue: the former may be concrete or abstract, the latter spatial, inferential, structural or causal (O'Brien & Opie 2004).

We lack a clear way of defining what counts as *bona fide* objects (and systems)[5]. For instance, if infiltration spots on a ceiling are to count as objects of the system surface-of-ceiling, it is trivial to find a partial second-order resemblance based on, *e.g.*, spatial relations between the spots and points on a city map. It is plausible that any two systems can be found to be in a second-order resemblance to each other, albeit the resemblance in most cases will be limited to small parts of the two systems and be somewhat *ad hoc*. The liberality of second-order resemblance underscores the main flaw of theories of content that use it as the content-fixing relation: wild non-uniqueness of content, a type of indeterminacy problem. I will come back to the issue of liberality and non-uniqueness of content in section §2.2.

[5]See Morgan (2014, p. 233) for an analogous point.

The notion of second-order resemblance is important because a theory of cognitive and intentional representation based on sharing first-order properties clearly does not get off the ground. For, rather obviously, the physical properties of neurons and assemblies of neurons do not resemble most of the properties that entities represented feature, such as their colours, textures, rigidity, shapes, and so on. My neurons are certainly not black, or smelly, and though they, just like penguins, cannot fly, they surely, in opposition to penguins, cannot swim. That does not preclude, of course, that neurons may represent penguins. From what we know from neurophysiology, moreover, this is a reassuring outcome: the physiological *substratum* that realises cognitive states, abilities and processes, among which representations, does not allow first-order resemblance in most cases, and even in the cases in which that kind of resemblance is possible (*e.g.* spatial relations between neurons mirroring spatial relations in the world), it is far from plausible.

Neurons operate in terms of excitatory and inhibitory relationships to each other, which are normally very complex. There is no conceptual obstacle — though there may be empirical ones — to the claim that neurons and assemblies of neurons might instantiate, through their excitatory and inhibitory connexions to each other, their patterns of activation, or some more complex property, aspects of the relational structure of entities in the world. Being a highly abstract relation, the sharing of relational structure is compatible with brain states standing in second-order resemblance relations to entities in the world. Due to its abstract nature, basing representation on second-order resemblance puts almost no constraints on the sorts of vehicles that can represent — the representation must only have a rich enough relational structure to mirror the relational structure of what is represented. Second-order resemblance is therefore a good candidate for the kind of relationship between representation and represented that can provide a plausible naturalistic resemblance theory of representation.

2.1.1 Structural resemblance and intrinsic properties

At the core of the notion of structural resemblance lies the idea that the structure which appears on one side of the resemblance relation is intrinsic to the system S_{V}, inasmuch as the relations in R_{V} are dependent on the physical properties of the elements in V. In other words, it is the intrinsic physical organisation of the system (or subsystem) that constitutes the relational structure that is shared with the second system. When inserted into the context of a theory of structural representation, the basic insight is that the intrinsic relational structure of a collection of representational vehicles — *i.e.* the physical relations between the vehicles — forms one of the *relata* of the (partly) content-determining resemblance relation to things in the world. Such a relational structure is intrinsic to the collection of vehicles inasmuch it is independent of any use to which such a collection may be put by the cognitive system.

Paradigmatic non-cognitive examples of structural representation are cartographic maps. In the case of cartographic maps, we have physical relations — distance between points — mirroring or resembling spatial relations between things in the world. Put

differently, when S_V is a map, the objects in V, namely points on the map, stand in spatial relations to each other which correspond to how (at least some) objects in O stand in spatial relations to each other. In the case of maps, spatial structure stands on both sides of the relation, but this need not be so. As Galileo's application of geometry to physics shows, spatial properties and relations of (abstract) geometrical figures can mirror relations between intervals of time, and the speed and acceleration of macroscopic moving bodies.

Structural representation requires that representations be part of structures. Even if the system of representational vehicles can be seen as composed of collections of basic unstructured representations, the latter can only be taken as representations derivatively, for they are so only insofar as they contribute to the structure of the whole system of vehicles, having no representational status when separated from such a system. As Cummins (1996, pp. 96-7) puts it, using somewhat different vocabulary:

> According to PTR [Picture Theory of Representation], there is no such thing as an unstructured representation, except in [a] derived sense ... an unstructured element [a vehicle] in a representing structure R [the collection of vehicles] may be said to represent its counterpart in a represented structure C. We must be careful, then, not to think of the objects, relations, and states of affairs in R as independent semantic constituents of R.

The idea behind structural resemblance is that it is physical relations between the representational vehicles in a representational structure that help establish the content-fixing second-order resemblance to things in the world. These physical relations are properties possessed by the representational structures independently of any use that the cognitive system may make of them — in this sense, they are intrinsic properties of the structured collections of representational vehicles. These intrinsic properties allow the vehicles to be individuated non-representationally — their physical relations are in place regardless of any representational use they may be put to by the cognitive system. It is those intrinsic properties that partially explain why those vehicles work as representations, since it is their physical relations that are supposed to ground structural resemblance.

Structural resemblance and structural representation offer a non-circular way of accounting for how physical structures in the cognitive system are able to play the role of representations. Their intrinsic relational structure helps explain how they come to stand in the relevant structural resemblance relation to things in the world, thus coming to represent them — and being able to be used as representations by the cognitive system.

This feature of structural resemblance, and thereby of structural representation sets it apart from related theories of representation that also appeal to second-order resemblance relations. Gallistel (1990), for instance, embraces second-order resemblance as the content-fixing relation, but proposes that it is partly the use that is made of the representational vehicles by the cognitive system that determines their relevant relational structure. By his lights, the relational structure of representational vehicles — one of

the *relata* in the content-fixing resemblance relation — is not intrinsic to the collection of vehicles; that is to say, it is not independent of the use to which those vehicles are put. In consequence, the relational structure of the vehicles cannot be the grounds for their being processed as they are, for their being suitably used as representations — making the appeal to second-order resemblance lose much of its explanatory power in accounting for successful behaviour.

In comparison to other theories of representation and content, structural representation as I have here presented it features a crucial advantage: it allows systems of representational vehicles to be partly defined by their intrinsic properties, providing a way to explain why those vehicles can be appropriately used as representations. It is because vehicles stand in the right physical relations to each other that the elements of cognitive structures stand in the content-fixing relation to things in the world. This explains, in a way that is independent of the representational use the vehicles are put to by the cognitive system, how they can suitably work as representations — as models or stand-ins for things in the world.

2.2 Structural Representation and non-uniqueness of content

Theories of content that rely mostly or exclusively on forms of second-order resemblance have to confront a difficult problem: resemblance relations are widespread and unconstrained. Basing content fixation on a liberal relation such as second-order resemblance makes representational content (wildly) non-unique. Given that second-order resemblance relations are common, structures stand in such relations to many different entities in the world. If second-order resemblance is at the basis of content-fixation, it follows that every representation represents many different entities in the world — representational content is wildly non-unique. Second-order resemblance theories of content thus risk falling prey to their own type of indeterminacy problem.

The so-called logical objections against resemblance theories of representation, moved by Goodman (1976), are a special case of this wider problem, to which structural representation is vulnerable. To see how damaging the problem is to structural representation, let us take as grounding content fixation a particularly strict form of sharing of relational structure: isomorphism. Though in the literature on structural representation — and on mapping-based theories of representation more generally — different types of morphism have been proposed as characterisations of (part of) the relevant representational relation, one of the most commonly appealed to is isomorphism [6].

Two structures V and O are isomorphic in a relation-preserving way if there is a one-to-one mapping such that every element of V corresponds to an element of O; every element of O corresponds to an element of V; every relation between elements of V corresponds to a relation between elements of O; and every relation between elements of O corresponds to a relation between elements of V. That is to say that the

[6] See, for instance, Cummins (1996), Millikan (2000, 2004), Waskan (2006), Ramsey (2007).

two structures completely share their relational structure, and their domains have the same cardinality. Second-order resemblance is here in its strongest form. Theories of representation that wish to employ relation-preserving isomorphisms to ground representational content have it that, for a structure to represent another, every element and every relation in the represented domain have a corresponding element and relation in the representing structure, and vice versa[7]. Most proponents of structural representation regard isomorphism as too strong a relation for grounding content-fixation, and settle for weaker forms of morphism[8].

Goodman's logical objections to resemblance theories apply to relation-preserving isomophism-based theories as well, as Suárez (2003) points out. Relation-preserving isomorphisms feature logical properties that are not shared by the representational relation. Isomorphism, as a mapping relation between two structures, is (a) reflexive, (b) symmetric, and (c) transitive. That is to say: (a) every structure is isomorphic to itself; (b) if a structure V is isomorphic to a structure O, structure O is also isomorphic to structure V; (c) if a structure V is isomorphic to a structure O, and structure O is isomorphic to structure U, then V is isomorphic to U. These properties follow from the definition of isomorphism. Any structure is structurally identical to itself; if a structure is structurally identical to another one, the latter will also be identical to the former; and analogously for transitiveness.

The representational relation does not generally seem to have any of these logical properties: (a) a representation R does not represent itself; (b) if R represents S, S need not represent R; (c) if R represents S, and S represents E, R need not represent E. These objections would show that isomorphism by itself cannot be at the basis of content fixation: other factors have to be invoked so as to block reflexiveness, symmetry and transitiveness. Otherwise, representations would have overly non-unique contents — *e.g.* a worldly state-of-affairs, themselves, etc.

More generally, isomorphism is too liberal a relation. Even relation-preserving isomorphism, which introduces a further dimension of complexity — *i.e.* there must be correspondence between relations among elements of the two structures — is too easy to come by, at least when it is left unconstrained. There is an isomorphism between any two sets of elements with the same cardinality. For if the sets have the same cardinality, a one-to-one mapping from elements of one to the elements of the other can be established. More importantly, as McLendon (1955) and Shea (2013*b*) point out, following the original insight by Newman (1928), unless the nature of the relational structure to be preserved in the mapping between two structures is constrained, there is always a relation-preserving isomorphism between two structures with the same number of elements.

To illustrate this point, McLendon (1955) uses the following example. Suppose that one finds twelve cars in a parking lot; each naturally standing in spatial relations to

[7]Cummins (1996, p. 96) brings this idea to its extreme, claiming that "the representational relation is just the relation of isomorphism".

[8]See, for instance, Swoyer (1991), O'Brien & Opie (2004), Waskan (2006), Bartels (2006), Shagrir (2012*c*).

each other. Suppose further that in coming home, one scatters twelve toy blocks on the floor. Given that the cardinality of the set of cars and that of the toy blocks is the same, there is an isomorphism between the two sets. More interestingly, there is at least one isomorphism between the two sets which preserves relational structure. For each spatial relation, whatever it may be, between cars in the parking lot can be made to correspond to a spatial relation, whatever it may be, between the toy blocks scattered on the floor (McLendon 1955, p. 90). For a theory of structural representation, it follows that any cognitive structure will stand in a relation-preserving isomorphism to any structure in the world that has the same number of elements. Cognitive structures would in consequence represent an indeterminate number of entities in the world, if standing in such a mapping relation is all that is needed to fix representational content.

Yet more damningly, McLendon (1955) presses on, without a principled way of individuating the elements of structures, the latter can always be decomposed such as to have a certain cardinality. Therefore, any two structures in the world can be so decomposed as to have the same number of elements. There would thereby be a relation-preserving isomorphism between any two structures. It follows from these considerations that any two structures in the world stand in a strict type of second-order resemblance relation to each other, namely isomorphism. Transposing these results to structural representation leads to triviality: any cognitive structure will stand in a content-fixing relation to any structure in the world, making representational content so widespread and non-unique as to lose any explanatory value. As Ramsey (2007, p. 93) puts it, understating somewhat the problem, if relation-preserving isomorphism is to ground the representational relation all by itself, then representational content will be indeterminate, any representation will be "potentially about a wide array of things".

This general argument for the liberality of relation-preserving isomorphism blocks one line of reply to Goodman's and Suárez's logical objections against resemblance theories of representation. While isomorphism is a reflexive, symmetric, and transitive mapping relation, weaker forms of relation-preserving mappings are not. For instance, Bartels (2006) appeals to a weakened form of homomorphism as the basis for the content-fixing relation. Homomorphism is a some-some mapping between sets and, as such, can take place between structures with different numbers of elements. Importantly, homomorphisms are neither symmetric nor transitive, though they are reflexive. On this account, a structure V represents a structure O when it respects two conditions: (a) every relation in V has a corresponding relation in O; (b) some relations in O have a corresponding relation in V. One may even weaken further the required relation and stick to the general notion of second-order resemblance, as O'Brien & Opie (2004) do.

The fatal shortcoming of this line of reply to the logical objections is clear. Though they deal with some of the issues raised by Goodman (1976) and Suárez (2003) by appealing to second-order resemblance relations that lack some logical properties that representations also lack, in so doing they fall into the clutches of the general argument above. For the second-order resemblance relations they appeal to are less strict than isomorphism, and as such, are even more widespread, and easier to come by — there

is not even the requirement that the two structures have the same cardinality. The triviality that follows for theories of content based on relation-preserving isomorphism *a fortiori* also applies to theories based on less demanding second-order resemblance relations.

In sum, structural representation, being based on overly liberal content-fixing relations — *i.e.* structural resemblance — suffers from a crucial defect: it makes representational content wildly non-unique, trivialising the import of the notion, and with it its explanatory value in the cognitive sciences. More generally, theories of representation that rely exclusively on second-order resemblance for content-fixation, such as the ones defended by Cummins (1996), O'Brien & Opie (2004), seem doomed.

At any rate, with structural representation we seem to have hit on a venerable, intuitive, and powerful idea about how representations work: they model the relational structure of things in the world, thereby standing in for those things in the operations of the cognitive system, allowing surrogative reasoning. This insight should not be dismissed lightly, for despite the failures of theories purely based on it, other paths to develop it into more satisfying theories of content are open. One such a path, deflationary in nature, will be the aim and theme of the bulk of this thesis. But before we move on to that, let us first briefly tackle sophisticated recent attempts to enrich structural representation with further content-determining factors so as to try and dodge indeterminacy problems.

2.3 Structural Representation: new hopes

Recently, some theorists have tried to remain faithful to the basic insight of second-order resemblance theories of content, while avoiding the wild non-uniqueness of representational content that follows from it (Bartels 2006, Ramsey 2007, Isaac 2012, Shea 2014). Their attempt is to curb non-uniqueness of content by adding further constraints on the factors that bestow representational content on cognitive structures. There are three main mutually non-exclusive strategies that can be followed[9]: a) add constraints on which cognitive structures are candidates for representational status; b) add constraints on which structures in the world are candidates for appearing in the contents of representations; c) narrow down the candidate content-fixing relation.

Let us briefly take a look at some of the forms these three liberality-curbing strategies may take.

It may be required that cognitive structures, if they are to be candidates for representational status, be in some way natural and non-trivial (Isaac 2012, Shea 2014): they must be non-arbitrary structures that are individuated by our best sciences of the mind and brain. Only structures that play a role in scientific explanation gain the status of relevant structures over which structural similarity can take place. This constraint keeps *ad hoc* and arbitrarily constructed groups of entities and relations in the cognitive system from being candidates for representational status, insofar as they are not legit-

[9]See Shea (2014).

imate, explanatorily useful groupings of entities for our best sciences — *e.g.* random selections of neurons and glia cells across different areas of the cortex.

Moreover, the elements and relations that characterise the cognitive structure must be such that they can be exploited by the cognitive system — that is to say, the cognitive system must be sensitive to them, and be able to use them in further processing (Shea 2014). Cognitive structures that stand in second-order resemblance relations to things in the world by means of properties that cannot be used in cognitive processing — simply because the system is not sensitive to those relations, such as neurons' colours — are thereby excluded from being candidates for representational status.

Constraints can also be imposed on the entities in the world that are candidates for being representational contents. Such entities may be required, for instance, to be in some way salient or significant for the organism (Shea 2014). Thereby, there might exist second-order resemblance relations between appropriate cognitive structures and entities in the world, but such relations will not help bestow representational status and content on those structures if the worldly entities at hand are not of the kind that can be salient or significant to a certain type of organism.

Finally, theorists have imposed further requirements on the structural resemblance relations that help bestow representational content on cognitive structures. Isaac (2012), and in a different way Bartels (2006)[10], propose that only those structural resemblance relations that have been generated by causal mechanisms linking the worldly entities represented to the representational vehicles count as content-bestowing — they add a causal component to the content-fixing relation. Not every second-order resemblance between appropriate cognitive structures and appropriate entities in the world bestow content on the former, but there must be some sort of causal relation between the two — causal relation that leads to the structural resemblance taking place.

In addition, Isaac (2012) requires that the representation be causally downstream from the entity represented. In sum, there must be a causal chain from the entity represented to the representation, and that causal chain is responsible for the existence of the structural resemblance relation between representational vehicle and content. A non-cognitive example is a footstep in the mud: the footstep represents the foot insofar as there is a structural resemblance relation between the two, and that relation has been established by a causal chain leading from the foot to the footstep, namely the pressing of the foot in the soft matter of the mud.

Another requirement on the content-fixing relation may be that the relations between the elements of the representational vehicle that mirror the relations between elements of the domain represented be projectable. This constraint helps to ensure that the relation-preserving structural resemblance between representational vehicle and entity represented is robust, reusable, and systematic — and not a chancy one-off coincidence. This preserves the stability of representations and their contents, as well as their use-

[10]Bartels (2006) appeals to causal relations only in order to pin down what representations are applied to — what he calls actual representation — while he embraces liberality for what he calls potential representations, which he equates with representational content. Therefore, Bartels ultimately accepts that content is wildly non-unique, but tries to save the explanatory value of representation by constraining representational application. A similar strategy is followed by Cummins (1996).

fulness for the organism in its dealings with the environment. Moreover, it avoids that randomly occurring structural resemblances between cognitive structures and entities in the world be counted as cases of representation.

Finally, some theorists have appealed to the use a cognitive structure is put to in order to pin down its determinate, unique representational content (Ramsey 2007, Shea 2014)[11]. The underlying idea is that an appropriate cognitive structure, standing in the appropriate structural resemblance relation to an appropriate entity in the world, is a representation of the latter only if the cognitive system employs it as such in the organism's interactions with the environment. For a cognitive structure to be a representation of a certain entity, in other words, the former must be used or exploited by the cognitive system as a representation of the latter, informing further processing and behaviour toward it. Adding use to the picture narrows down considerably the liberality of second-order resemblance, since only when cognitive structures are actually used by the cognitive system due to their standing in second-order resemblance relations to entities in the world do they gain representational status and content.

In sum, there are several mutually consistent strategies that can be employed in trying to save structural representation from the liberality of structural resemblance, thus keeping at bay, at least to some extent, the non-uniqueness of representational content that follows. There are though reasons to be at least moderately sceptical about whether these strategies suffice to make structural representation a satisfactory theory of content.

Firstly, it is not clear that the proposed constraints, even when combined, are able to avoid some degree of pernicious liberality of representation and non-uniqueness of content. For instance, the appeal to exploitability is certainly not enough, since one and the same exploitable cognitive structure can be used as a representation of anything that it structurally resembles. On the other hand, strengthening the requirement so as to include only actually exploited cognitive structures risks closing the doors to the explanatorily useful notion of unexploited content (Cummins et al. 2010/2006).

Secondly, some of the constraints can be difficult to justify in a principled way. For instance, being part of the current explanatory practices of our best sciences is a criterion that cannot be easily cashed out, given the variety of scientific approaches and the ways in which they carve cognitive systems; and it seems moreover to be too depend on current knowledge and practices to be able to help ground a general theory of cognitive representation. Similarly, the appeal to projectability enmeshes debates about representation with complex and disputed issues in general philosophy of science.

Thirdly, there is a risk that the factors added to structural resemblance will introduce further indeterminacy problems in the account. For instance, the appeal to causal relations made by Isaac (2012), while considerably limiting the non-uniqueness of content stemming from exclusive reliance on resemblance relations, launches the resulting account into the throes of the problem of error and the distality problem. Moreover, as Isaac (2012, p. 701) admits, his theory of structural representation does not make

[11] I will come back to Ramsey's view with more care in section §6.3.

space for misrepresentation: every cognitive structure standing in a causally-generated structural resemblance relation to the world represents correctly whatever entities in the world caused that resemblance relation to take place. As for causal-informational semantics, no straightforward account of the normativity of content is on hand.

Similar worries infect theories of structural representation that appeal to the use to which representations are put by the cognitive system. How to distinguish correct from incorrect uses? If, as it seems more plausible, this is to be effected by having recourse to some form of teleology, issues of functional indeterminacy become pressing. Moreover, the Cummins/Burge argument against equating representational and behavioural success may also apply.

In short, though the prospects for theories of representational content that rely on forms of second-order resemblance are not as bleak as the non-uniqueness of content problem might initially suggest, the proposed fixes risk to close one hole in the hose while opening new ones elsewhere. The further factors appealed to in curbing liberality and non-uniqueness of content open up other indeterminacy issues typical of mainstream theories of content, such as causal-informational semantics and teleosemantics.

2.4 Looking elsewhere: a deflationary approach to representation

After this brief overview of the main options in the debate about representation and content, as well as of the main challenges they face, it is time to move to my own take on these issues. My deflationary proposal, I take, is an useful addition to this debate, one that suggests that a different path can be treaded in accounting for representation in the cognitive sciences. A path that sees representation and content as less metaphysically-loaded notions; and as a more flexible, context-sensitive, and fluid affair than what the mainstream proposals would have wanted. This approach may also help dissolve some of the traditional issues that have exercised theorists interested in representation and content, in particular indeterminacy problems.

The hunt for a satisfying deflationary theory of representation and content for the cognitive sciences has in Part II its prelude, whilst Part III will see its development and (hopefully) happy epilogue. My strategy will be to focus first on representation's sister foundational notion in the cognitive sciences — computation. I claim that a non-trivial, non-liberal, robust notion of concrete computation, developed in Part II, provides the grounds for unloading part of the explanatory burden normally placed over the notion of representation, opening thus the doors to deflating the latter. Structural representation, after remaining behind the curtains for the whole Part II, will make a momentous reappearance, though in a different guise, in Part III. I aim to show how, suitably deflated and freed from robust reductionist scruples, structural representation provides an important share of the story about the nature and role of representation and content in the cognitive sciences.

With the lay of the land under our eyes, it is time to proceed to the positive side of

the project.

Part II

Computation and Mechanism

Chapter 3

Concrete Computation

That representation came to play, since the beginning, a foundational role in the growing and now maturing cognitive sciences is not surprising, or revolutionary. The idea that cognition involves representation is an old one, and its roots can be traced back at least to Plato. The birth and blossoming of the cognitive sciences as a structured research field owed its keep primarily to the 'perfectly stunning idea'[1] that cognition worked in a fashion fundamentally similar to that of the then dawning computers, whose existence was made possible by mathematical developments in the first half of the last century, and which saw the likes of Turing, Church, and Von Neumann as protagonists. Underlying cognitive phenomena, the proposal was and is, are computational processes that 'manipulate' — transition between or transform — representational states according to rules.

Taking the cognitive system to be a computational system has several advantages. Theorists studying cognition can use models and tools developed by the mathematical theory of computation. Moreover, the availability of computers provides a ready way to try and test hypotheses on how a cognitive task can be solved by computational means. But most importantly, appeal to computation promises to give a way of understanding how appropriate and rational behaviour is possible in physical systems.

The rules that determine the behaviour of a computational system can explain how the transitions between representational states come to be adequate to the task at hand. By employing the right rules, computational systems can solve any computable problem. If the cognitive system is computational, we have a ready way to understand how goings-on in the brain can lead the organism to adequate behaviour towards the body and the environment: the cognitive system is making use of the suitable set of rules, or the right 'programme' for the situation at hand. The brain can thus be seen as the hardware in which the computations carried out by the cognitive system are implemented, in analogy to the silicon chips that implement the computations and programmes that govern the behaviour of modern electronic computers. Thence the famous idea that the mind is the software of the brain, *i.e.* the computations implemented, and programmes stored, in the brain.

[1] The expression is Jerry Fodor's, referring to an idea he traces back to Alan Turing — and which had precursors, such as Hobbes.

Furthermore, appeal to computation dispels the mystery of how cognition is possible in physical systems. By means of a theory of computational implementation, the computations that govern adequate behaviour can be explained in purely physical terms by showing how the relevant computational processes are realised by the biochemical machinery of the brain. An underlying causal story can thus be told of how the cognitive system is able to sport complex adequate behaviour in its day-to-day interactions with the world.

Finally, placing computation at the foundations of the cognitive sciences ties nicely with its foundational colleague, the notion of representation. At least in the philosophical literature on cognitive science, computation is widely seen as essentially involving the manipulation of states endowed with representational content, so much so that Piccinini (2008*a*) and Sprevak (2010) dub it 'the received view about computation'[2]. The idea, expressed synthetically in Fodor's famous slogan — 'No computation without representation!' — is that computation only takes place when there is trafficking in states with content. On this common understanding, representation, computation, and cognition are intimately tied together: taking computation essentially to involve contentful states builds a straightforward bridge between that notion and the other two, being thus particularly fit for the purposes of the cognitive sciences.

However, as for the notion of representation, several conceptual issues need to be addressed if computation is to play its foundational role in the cognitive sciences appropriately. While in mathematics the notion is abstract, and independent of any physical substrate, the needs of cognitive science require an account of concrete computation, or computation realised in physical systems. In particular, two questions need to be answered: what makes something into a computation?; what conditions must hold true of a physical system so that it performs computations?

The former question is about computational individuation, while the second is a question about computational implementation. The two are closely related: if a certain property Z plays a role in individuating a computation, then any physical system that implements that computation must also have property Z. If it failed to do so, it could not count as performing that computation, or perhaps any computation at all. Therefore, issues of individuation and implementation are closely knit together when it comes to concrete computation.

Satisfactory answers to these questions are crucial to ensure firm conceptual grounds for the cognitive sciences, at least for what regards one of its sustaining pillars. As Sprevak (2012) points out, a good theory of concrete computation should meet at least the following three *desiderata*[3].

First, (i) it should clarify the notion of concrete computation, often treated as an

[2]Interestingly, there is controversy even for what regards what the most common view is. Shagrir (1999), for instance, holds that the standard view of computation is the non-semantic one. I believe that this disagreement stems from the scientific field one examines when making such claims. The semantic view is more common in philosophy of mind and cognitive science, while the non-semantic view in computer science. See O'Brien (2011) for similar considerations.

[3]Piccinini (2007*b*), Fresco (2014), Piccinini (2015) propose six *desiderata*. For my purposes here, these three (plus one) will suffice.

explanatory primitive. Second, (ii) it should help vindicate the specialness of computational explanation by driving a wedge between systems that actually implement computations, and systems that do not. Third, (iii) it should provide an account of concrete computation in which facts about physical systems implementing computations are objective features of the world, independent of the intentions and beliefs of observers. The objectivity of concrete computation is particularly crucial for its foundational role in the cognitive sciences. As Ladyman (2009) points out, ascribing computational processes to the cognitive system only helps naturalising cognitive phenomena if performing computations is a natural fact, independent of the ways human beings represent the world. As a fourth *desideratum*, I add normativity: theories of concrete computation should do justice to the normative aspect of computation, *i.e.* that there be something it is for a computation to be correctly or incorrectly performed — that there can be instances of miscomputation[4].

Several theories of concrete computation have been offered, and I will briefly examine some of them in the following sections. As O'Brien (2011) notes, such theories fall on two opposing sides of a fracture[5]. On each side a different criterion for computational individuation is proposed. The first one, 'the received view' already hinted at above, has it that computation essentially involves representation, and that computations are at least partially individuated by their semantic properties. This line is generally dubbed 'the semantic view'. On the opposing side there are those who argue that computation need not involve contentful states, and that computations are individuated non-semantically — the 'non-semantic view' of computational individuation. Each side of the fracture features its own advantages and shortcomings, and it is the aim of this and the following chapters to bring them forward and provide an overview of the conceptual space for theories of concrete computation.

I will start with non-semantic proposals. First, I will present and examine the simplest account of concrete computation, the mapping account. I will then examine its chief problem: it makes computation trivial, leading to unlimited pancomputationalism. There are three main strategies to cope with the triviality objections against the mapping account: accept that concrete computation is trivial, and recommend the ejection of computation from the foundations of cognitive science (or of any other science, for that matter); accept triviality, but try and save the explanatory purchase of the notion; or supplement the view with further constraints so as to block triviality arguments.

In this chapter, I will explore these three strategies. In section §3.1, I present the simple mapping account, and the trivialisation problems that follow from it. After that, I will, in section §3.2, examine a pragmatist version of the simple mapping account that accepts triviality, while nevertheless trying to defend the explanatory value of computation to the cognitive sciences. Finally, section §3.3 examines the advantages and shortcomings of two sophisticated families of views of concrete computation. In section 3.3.1, I analyse a refined non-semantic view, based on causal mapping, put

[4]See Fresco & Primiero (2013), Piccinini (2015).
[5]An exception to this may be Rescorla (2012*b*).

forward most forcefully by Chalmers (2011). In its turn, section 3.3.2 presents and assesses the main arguments in favour of adding a semantic factor to theories of concrete computation. I will argue that most of those arguments fail, but one — the argument from the multiplicity of computations — spells trouble for non-semantic views. The following chapter will put forward a theory of non-semantic concrete computation, the mechanistic one, that, I argue, has the tools to avoid that argument.

3.1 Mapping accounts and the Putnam-Searle triviality objections

Mapping accounts claim that a physical system implements a computation when there is a mapping between the formal structure of the computation and the physical states and processes of the physical system. Given an abstract computational description, a physical system implements it if its states and state-transitions as determined by a physical description mirror the states and state-transitions defined by the abstract computational description[6]. This view is associated with Putnam (1988).

This intuitive definition of the view invites an intuitive objection. Take any complex physical system, such as a wall or a pail of water. Take any computational description, such as the one specifying the programme LyX, the word editor used to write these pages. The wall, or the pail of water, are composed of a prodigious number of atoms and subatomic particles in different types of interaction, transitioning from physical state to physical state as time passes. It seems very likely that at least one of the patterns of state-transitions taking place in the wall or in the pail of water mirrors the state-transitions specified by the programme LyX, at least within a certain time interval. It would follow that the wall, or the pail of water implement the programme LyX as much as my laptop computer.

Searle (1992), developing insights by Ian Hinckfuss, claims that, given the enormous amount of patterns of microphysical transitions that take place in any complex enough system, it is extremely likely that a mapping can be found between at least one of such patterns and the states, and state-transitions specified by any computational description. Therefore, according to Searle, computation is trivial insofar as any physical system complex enough implements any computation*:* the mapping view leads to unlimited pancomputationalism[7].

If concrete computation is trivial, cognitive science is in deep trouble. For it follows that appealing to computation in explaining cognition is vacuous. Any complex system, Searle's triviality argument purports to show, performs computations; not only designed computers and cognitive systems. Taking the cognitive system to perform computations does not set it apart from any other complex system in the universe, including walls and rocks. Moreover, the brain itself, as a complex system in its own right, implements every computation. The basic idea underlying cognitive science, *i.e.* that we can ex-

[6] As Piccinini (2015, p. 17) points out, it is more precise to talk of 'microphysical states', insofar as computational descriptions are also physical.

[7] Piccinini (2015).

plain cognitive processing and complex behaviour by means of finding out the specific computations implemented by the cognitive system, would founder.

Searle's intuitive argument applies to the rough characterisation of mapping-based computational implementation provided above. It is worthwhile to flesh out the picture more carefully. In order to provide a detailed picture, mapping views have to define the two structures that enter into the mapping relation. One is the structure of the implementing system given by the physical description, consisting of its physical states and state-transitions. The other is the formal structure of the computation as determined by its computational description. This brings to the fore the importance of computational formalism in mapping accounts. Different computational formalisms lead to different descriptions of the abstract structure of a computation. Therefore, they characterise differently one of the two *relata* of the mapping relation that establishes computational implementation. The appropriate formalism should be chosen according to how adequately it fits the needs of computational theorising in the cognitive sciences (Chalmers 1996).

Putnam focuses on finite-state automata (FSAs). An FSA is an abstract computational formalism defined by the states an automaton can be in (only one at any particular moment of time), and a table of state-transitions. Inputs and outputs can also be added to the formalism. Any FSA can be only in a finite number of different states. An FSA can thus be specified by providing the set of its internal states $\{S_1,..., S_n\}$, the set of input $\{I_1,..., I_n\}$ and output $\{O_1,..., O_n\}$ states, as well as rules that specify the state-transitions from each pair of internal state and input state to each pair of internal state and output state, S_i , $I_i \rightarrow S_j$,O_j. Examples of implemented FSAs are automated coffee distributors and lift controllers.

Given the FSA formalism, it is possible to define the mapping account of concrete computation more fully. A physical system implements an FSA if and only if

> there is a mapping M from states of the physical system onto states of the FSA, and from inputs and outputs of the physical system onto inputs and outputs of the FSA, such that: for every state-transition $(S, I) \rightarrow (S', O)$ of the FSA, if the physical system is in state P and receives input I^* such that $M(P) = S$ and $M(I^*) = I$, then it transitions to state P' and emits output O^* such that $M(P') = S'$ and $M(O^*) = O$.[8]

Putnam (1988) shows, however, that the FSA-based mapping account of computational implementation leads to triviality of concrete computation. The strongest result comes from analysing inputless FSAs, that is, finite-state automata without input or output. These automata transition between internal states following rules of the form $S_i \rightarrow S_j$. Putnam proves that any arbitrary open physical system implements any inputless FSA. The only assumption required for his proof to go through is that the physical system have non-cyclical behaviour, that is, that it always be in different internal states at different moments in time[9].

[8] Adapted from Godfrey-Smith (2009*b*).

[9] Chrisley (1995) argues against this assumption. Most theorists, in contrast, accept that the as-

Here is a simplified exposition of the proof. Take any open system T in a specific time interval, say from 12.00 to 12.07. Group as belonging to state S_i of T the section of the trajectory in state-space of the total physical state of the system during each minute from 12.00 to 12.07. Hence, in that time interval the system will transition between states S_1, S_2, ..., S_7. Now take an inputless FSA that undergoes the following transitions in that same interval, $A \rightarrow B \rightarrow A \rightarrow B \rightarrow A \rightarrow B \rightarrow A$. Define physical states a, b of T, $a = S_1 \bigcup S_3 \bigcup S_5 \bigcup S_7$ and $b = S_2 \bigcup S_4 \bigcup S_6$, and map a into A and b into B[10]. The physical system undergoes the same state-transitions as the FSA in that time interval. Therefore, according to the mapping account, it implements that FSA during that time interval. The proof generalises to any (non-cyclical) open physical system, and any inputless FSA. Therefore, any physical system implements any inputless FSA, making concrete computation trivial, and leading to unlimited pancomputationalism.

A similar, though weaker result, applies to FSAs with inputs and outputs. An analogous proof shows that any physical system that has inputs and outputs mappable to the abstract inputs and outputs of an FSA will implement it. So, provided there is a mapping between the inputs and outputs of the physical system, and the inputs and outputs of the computational description, a physical system will implement every FSA with the same input-output function. The mapping of the input-output function is open to two interpretations: a weak, and a strong one.[11].

The weak interpretation requires only that there be a mapping between the inputs and outputs of the physical system, and those of the FSA, without placing constraints on the physical nature of the inputs and outputs of the former. On this reading, the triviality of implementing FSAs with input and output is as strong as that of implementing inputless FSAs. The appropriate inputs and outputs in the physical system can be appropriately picked out from an arbitrarily chosen boundary using the same disjunctive strategy as above. Finding physical systems with (weak) input-output functions mappable into the input-output function of an FSA is again a trivial business.

The strong interpretation requires that the mappable inputs and outputs be of a specific physical type. The implementation of the LyX word editor must take key strokes as input, and display characters on a screen as output. A weaker triviality result applies to this strong interpretation. Any physical system with the appropriate strong input-output behaviour implements every FSA with that input-output behaviour. This means that any physical system that implements the LyX word editor likely also implements LibreOffice Writer, and Microsoft Word[12].

In sum, even in the case of FSAs with inputs and outputs pernicious triviality of concrete computation follows. All that matters for implementation, in this case, is having the right inputs and outputs. The internal computational structure of a system is irrelevant. In the case of cognitive science, Putnam (1988) points out, computationalism and behaviourism become indistinguishable.

---

sumption is plausible at least for macroscopic open physical systems, as do I.

[10]Chalmers (1996) adds this step to Putnam's proof for more precision.

[11]Sprevak (2012, pp. 120-1).

[12]Without regimentation, this would be true at least to a large extent.

In face of these results, three lines of reaction are available: (a) accept that computation is trivial and devoid of explanatory power (Putnam and Searle, at least for what regards cognitive science); (b) accept that computation is trivial, but defend that nonetheless it has explanatory purchase (Egan 2012, Schweizer 2014, 2016); (c) take the triviality arguments as a demonstration that the simple mapping account of concrete computation[13] presented above is wrong or incomplete, and provide corrections or additional conditions for computational implementation (almost everyone else).

Line (a) is particularly unattractive. It is to withdraw prematurely from a game with quite high stakes. It fails to do justice to efforts in computer science and electronic engineering — conceiving and building computers is far from a trivial matter — and it shakes the foundations of those sciences as well as of cognitive science. It becomes senseless to talk of different computational systems having different computational powers, since nearly any physical system computes nearly anything. The unlimited pancomputationalism that the simple mapping account leads to — every complex system performs every computation (or almost) — puts in danger the objective status of computation. If everything computes nearly everything, it would seem, nothing really computes anything — talk of systems performing computations is just vacuous.

Taking line (a) does not affect the mathematical theory of computation, for whose abstract objects the problem of implementation does not arise. But it deals a fatal blow to those scientific fields that see claims about physical systems performing computations as meaningful and explanatorily powerful. Line (a) recommends ejecting the notion of computation from any valuable explanation of the behaviour of physical systems. However, given how much is to be lost, how many promising scientific fields would have to be rethought, it is too untimely a retreat. Rather, (a) should stay as the last (apocalyptic) resort, the position to take when all other possibilities have been exhausted, and every hope is lost.

In the next section, I will analyse (b). In the following section, I will turn to (c), which will introduce the semantic *vs.* non-semantic divide between theories of concrete computation, and lead the way to my main focus in this part: present and defend the mechanistic view, a task for the next two chapters.

3.2 A pragmatic take on implementation

There may be a way to save the explanatory power of concrete computation, at least for what regards the cognitive sciences, while at the same time accepting that computational implementation is trivial. Something along the lines of this position is upheld by Egan (1999, 2009, 2012), Schweizer (2014, 2016), Matthews & Dresner (2016) and also contemplated by Chrisley (1995)[14], while Searle (1992) may be seen as taking some steps in this direction before veering toward a different path. The view involves appealing to

[13] The baptism is due to Godfrey-Smith (2009*b*).

[14] Chrisley (1995) does not hold this position, as he believes the Putnam-Searle triviality objections to be wrong-headed. In any case, he argues that even if the objections should go through, they would not suffice to deprive appeal to computation of its explanatory powers.

our scientific practices.

Research in the cognitive sciences normally takes as its starting point a cognitive capacity sported by an organism. The cognitive capacity is the target of the explanation, it is given, assumed from the get-go. By means of theorising and experimenting, cognitive scientists investigate how the cognitive system enables the feat. Computational cognitive science proceeds by trying to find out the computations performed by the system that explain the capacity.

Take, for instance, the ability of desert ants to go back to their home colony in a nearly straight line. The capacity is particularly striking given the windy path they generally take when foraging outside the nest, and the fact that they do not rely on landmarks[15]. A computational cognitive scientist interested in understanding how this sophisticated behaviour is made possible will look for a computational procedure, an algorithm, that would allow the ant's brain to calculate the straight trajectory, perhaps by integrating information on heading and average speed during the foraging journey (path integration).

Suppose that the cognitive scientist individuates a neural circuit in the ant's brain that implements that algorithm, *i.e.* the states and state-transitions of the circuit map onto the steps of the algorithm. They would then have come up with a satisfying computational explanation for the path integration capacity of the desert ant. Even though the ant's brain implements nearly every computation (courtesy of the triviality arguments), the computations that explain its navigational ability are the ones individuated by the scientist. Out of the indeterminate number of computations implemented by the ant's brain, that particular one explains the cognitive capacity under investigation.

As should be clear, this is too fast. Given the triviality of implementation, cognitive scientists would not have needed to look inside the brain of the ant to find a circuit that implemented the hypothesised algorithm. That the ant's brain implements that algorithm, as well as an indeterminate number of others that effect path integration, is obvious. After all, as a complex system, it implements nearly every algorithm! It seems that we ended up with the discouraging conclusion drawn by Putnam and Searle. From the triviality of concrete computation follows that computational explanation in the cognitive sciences is senseless.

However, appeal to scientific practice introduces further constraints on what implementations are explanatory. First, we know from the outset what the input-output function is. This limits the possible implementations of the computation to those physical systems that have mappable input-output functions. This limitation, if interpreted weakly, is of little consolation, as we have seen above. Every physical subsystem of the ant's brain with the same input-output function implements the algorithm. And it is trivial to pick out groupings of 'boundary' physical states that appropriately map onto that function. If, on the other hand, the input-output mapping is taken strongly, we

[15]Gallistel (1990). It is controversial whether this behaviour counts as cognitive. As is well known, there is no principled boundary between cognitive and non-cognitive, as we still lack any consensus on the 'mark of the cognitive' (or on the existence of such a thing, to start with). See Adams (2010). At any rate, this point is immaterial to my purposes here.

do make some headway. The inputs must come from specific sensory systems and the outputs must consist in activities of the motor system of the animal[16].

Second, and consequently, we know, or we can find out, which sensory systems convey the information needed to feed the steps of the algorithm. The circuit that implements the algorithm has to be causally connected both with the sensory channels that pick up and transmit information about heading and speed, and with the effectors that make the ant move. All implementations of the algorithm that are not appropriately causally connected to the relevant transducers and effectors cannot explain how the ant manages to get back to the colony in an approximately straight line. This puts a strong constraint on which circuits are implementing the explanatorily relevant algorithm[17].

Third, science tells us what the relevant causal level is, thus ruling out arbitrary groupings of physical states. One of the main reasons for the success of the Putnam-Searle triviality arguments stems from the lack of constraints on how to carve up and group the implementing physical states and processes[18]. If any grouping of microphysical states of a physical system counts as a candidate for implementation, triviality ensues. Indeed, Putnam's proof depends on a grouping of maximal physical states based on arbitrary temporal intervals.

Cognitive science individuates, to some approximation, the causal level that is relevant for its explanatory purposes. The causal goings-on of relevance are the ones involving sensory transducers, neurons, groups of neurons, and motor effectors. The behaviour of sub-atomic particles, atoms and so on, is not relevant to cognitive explanation[19]. Moreover, groupings of physical states that ignore boundaries between the entities posited by the cognitive sciences, *e.g.* that group together as belonging to one state arbitrary parts of different neurons and glia, are ruled out.

In sum, appeal to the practices and theoretical posits of cognitive science significantly limits the computational implementations relevant to its explanatory projects. Even though the cognitive system, as any complex system, performs nearly every computation, the implemented computations that explain the target cognitive capacities are severely constrained. I will refer to this view, with some terminological hesitation, as 'limited pragmatism'.

Is limited pragmatism about computational implementation suitable to save the explanatory role of computation in cognitive science? There are, I take, two main reasons for scepticism.

Firstly, it is unclear whether the view is able, by having recourse to scientific practice, to single out the one concrete computation performed by the cognitive system that does the explanatory job. Recall the weaker triviality argument that applied to FSAs with strongly interpreted inputs and outputs. It seems that that pernicious result applies to the foregoing account: even appealing to the practices and posits of cognitive

[16]See Schweizer (2014).

[17]See Egan (1999, 2012), Schweizer (2014).

[18]Scheutz (2012).

[19]See Egan (2012, pp. 46-7), who presses a similar argument based on the cognitive exploitability of properties: only macro-properties can be cognitively exploited in processes that perform computations relevant for accomplishing cognitive tasks.

science, we are still left with little more than constraints on the appropriate inputs and outputs. It follows that the cognitive system implements every algorithm with the same inputs and outputs. Cognitive science would hence not be able to distinguish different computational structures at play in bringing about cognition, collapsing, as Putnam warned, into behaviourism.

There may be a way out of this objection, even though, I believe, it is going partially to depend on empirical considerations. For the pragmatist has an additional card in their sleeve, as we have seen above — the explanatorily relevant causal level of the system is approximately fixed. The answer to the problem lies on whether the physical goings-on in the neural circuits which stand in the appropriate causal relations to inputs and outputs could be mapped onto different algorithms, as the simple mapping account of implementation would want; and on whether those different algorithms would be explanatory of the cognitive capacity under investigation.

Ladyman (2009), Sprevak (2010), and Shagrir (1999, 2001, 2012*a*,*b*) argue that the same causal structure of a system can implement two syntactic structures at the same time. *A fortiori*, this would mean that any circuit individuated by the foregoing view, even if unique, would perform at least two different computations. I will leave discussion of Sprevak's and Shagrir's argument to section 3.3.2 and section §4.4, but suffice it to say for now that their argument does not directly impinge on the foregoing account. They would need to show not only that the same neural circuit performs two different computations, but also that both computations are equally explanatory of the cognitive behaviour under investigation[20]. They would have to be computations of two algorithms appropriate to the same cognitive task. Shagrir's argument does not show that[21].

A second problem regards the fact that limited pragmatism involves adding an observer-dependent factor into the account of implementation. Concrete computation is objective, but trivial. Explanatory computations, the ones of use to cognitive science, are, on the other hand, dependent on the interests and practices of cognitive scientists themselves. Mind-dependent properties infect the account to some extent. Concrete computation keeps its explanatory power despite its triviality, but the explanatory purchase of the notion is rescued by appeal to intentions and subjective goals. The latter play a considerable role in determining which concrete computation, out of the indefinite number of them that each physical system implements, is explanatorily relevant in each case.

There is a risk, therefore, that the account will lead to vicious circularity when computation is seen as one of the foundations of the cognitive sciences: we want to explain cognition partially by means of concrete computation, but we use cognitive states to help determine which concrete computations are explanatory. The *explanandum* seems

[20]See Fresco (2015).

[21]Nor does it intend to do so. His target is rather non-semantic views of concrete computation, especially when conjoined with the claim that performing certain types of computations is sufficient for having a mind (computational sufficiency thesis). I will come back to this argument in the next chapter, as it threatens the mechanistic view of computation with computational indeterminacy, as it does the causal mapping view. The latter, though, is happy to bite the bullet from the get-go, as we will see below.

to be explaining the *explanans*, as well as the other way around.

One way to tackle such an accusation is claiming that, after all, the objectivity of concrete computation is preserved, though it be a vacuous notion. The appeal to interest-relative properties only comes in when one concrete computation, out of the infinite ones, is selected as explanatory of a capacity of interest to our scientific endeavours. If computation is interest-relative in the limited sense that explanations pick out one computation from the various ones performed by the system, based on the aims and targets of theorists, nothing detracts from its objectivity[22]. It is unclear, however, whether this would help save the foundational role of computation in the cognitive sciences. For computation, though objective, would still be a trivial matter. Unfortunately, I cannot discuss this strategy in any detail here, as it would require touching on some fundamental issues in general philosophy of science regarding the nature of explanation, and the relationships between instrumentalism and realism about scientific posits.

A second way to tackle the point is to accept that computation cannot play any foundational role in cognitive science, but maintain nonetheless that computation is a useful scientific notion. This strategy involves going full-out pragmatist, and holding that computation is always observer-dependent. This is stronger than the claim we have been analysing so far, namely that computation is observer-independent but trivial, and that pragmatic considerations come in to pick out explanatory computations from those that are not. On the foregoing, in contrast, computation is not objective to start with. In consequence, computation cannot lie at the foundations of cognitive science because it would lead to the vicious circle above. I will call this view 'full pragmatism'.

Searle's (1992) trivialisation arguments are supposed to support the view that the notion of computation is observer-dependent, a view also endorsed by Schweizer (2014). Whether a system is computing and what it is computing, for Searle, depends on observers interested in ascribing specific computations to systems. Schweizer (2014) offers a more thorough and well-argued defense of full pragmatism. He argues that pragmatic constraints help distinguish cases in which seeing physical systems as performing computations is scientifically useful and fruitful from cases in which it is not. Such constraints are moreover not rigid and unchangeable, but are rather dependent on the context — different explanatory and practical contexts will motivate the employment of pragmatic constraints which may play no role in other contexts[23].

We talk about physical systems performing computations because in some cases this is useful for our purposes, be them in explanation or in engineering. The overwhelming majority of possible computational interpretations that follow from the simple mapping account are useless and uninteresting. These are not in competition with the few, if any, computational interpretations that advance our understanding of the workings of physical systems, of how to predict their behaviour, as well as of how to regiment

[22]See Chrisley (1995). Moreover, if conjoined with an ontic view of explanation, the view could arguably fully shed its interest-relative, pragmatist, commitments, and embrace objectivism about explanatory computation. I just point at this possibility, which I will not investigate here.

[23]Schweizer (2016).

physical systems to causally behave in the ways that we want. Therefore, the fact that any complex enough physical system can be interpreted as performing any computation is irrelevant, for very few computational interpretations turn out to be scientifically and practically useful, and in line with what are considered to be good explanations.

Computers are those physical systems whose computationally relevant states we can discriminate, and that are able to carry out algorithms we are interested in over many different input values. Further pragmatic constraints on physical systems that can be helpfully seen as computers, according to Schweizer (2014), include: automaticity, reliability, versatility, and predictive power. The latter constraint rules out, as pragmatically irrelevant, those implementations that are arrived at only *ex post facto*, as in the case of Searle's wall, or the pail of water. There is no predictive power gained in assigning computations to those physical systems — quite on the contrary, it is only after observing their development during an interval of time that a computational interpretation can be offered, thereby adding nothing to our capacity of predicting how the systems will behave next, and rendering them useless as devices for performing computations.

Note that the pragmatic constraints that Schweizer proposes are in principle compatible with limited pragmatism, that is, a theory in which the objectivity of computation is preserved, but pragmatic constraints come in only in order to tell apart the explanatory computations from the non-explanatory ones. However, Schweizer clearly wants to subscribe to the full pragmatist view when he claims that he supports "the conclusion that realising or implementing an abstract computational procedure is not an intrinsic property of physical systems, but rather is based on a purely observer-dependent act of ascription"[24]. It is arguable whether Schweizer's arguments provide any reason to prefer full pragmatism to limited pragmatism. On the contrary, given that limited pragmatism preserves the observer-independent nature of concrete computations, while at the same time having the same consequences regarding the explanatory purchase of computation as full pragmatism, I believe that it is superior to the latter.

It is moreover difficult to assess whether computational explanation, by full pragmatist lights, would have any purchase, and whether the notion of computation would be able to play a foundational role in the cognitive sciences[25]. Again, the issue depends on one's own overall stand on realism, anti-realism, and instrumentalism about scientific posits, and their explanatory power. I will not go into these problems here, but instead I will assume, with, I take, most of the literature, that only observer-independent entities and processes are genuinely explanatory for most of the fields composing cognitive science. This stance naturally rules out Searle's and Schweizer's observer-relative, fully pragmatic notion of computation from playing any robust explanatory role in such fields; while it leaves unscathed the limitedly pragmatic view which insists that concrete

[24]Schweizer (2014).

[25]Churchland et al. (1990), who take this path and accept the observer-relativity of concrete computation, defend that it is explanatorily valuable nonetheless. On the other hand, Ramsey (2007, pp. 100-2) claims that deeming computation to be observer-relative undermines the appeal of the notion in computer and cognitive science. For Searle, the observer-relativity of computation supports the latter view, and is meant to be an objection to computationalism in the cognitive sciences.

computation is an observer-independent, though trivial, feature of the world.

It is worth keeping in mind that, if the robust, non-trivial theories of concrete computation that I will examine below should fail, there is still an arguably satisfying way to justify the use of the notion of concrete computation in our scientific and engineering practices; a fallback position that avoids most of the shortcomings that plague a pure simple mapping account: limited pragmatism. Such a view may involve embracing the triviality of concrete computation, while at the same time insisting that pragmatic considerations preserve its explanatory power in cognitive and computer science.

3.3 No computation without representation?

Most of the theorists interested in concrete computation have opted for strategy (c), namely complementing the simple mapping account with further constraints so as to block trivialisation arguments, and save the explanatory power (and the objectivity) of concrete computation. These constraints can be divided into two general categories.

The first category comprises constraints that have to do with semantic properties. Theories that make use of such properties — semantic theories — add to theories of concrete computation the requirement that computational states be representational[26]. For proponents of this view, computations are individuated partly by their semantic properties, and thereby in order to implement a computation the implementing system must have the adequate semantic properties — computation essentially involves representation.

The second category comprises constraints that do not involve semantic properties. Theories of concrete computation that appeal exclusively to non-semantic properties are variegated, insofar as different non-semantic considerations may be brought to bear in determining the computational nature of physical systems. The further factors appealed to include counterfactual support[27], dispositional properties[28], causal organisation[29], and functional properties[30]. In what follows, I will focus especially on the latter two.

Most semantic theories incorporate some of the further constraints put forward by non-semantic theories[31]. They then add the further requirement that a system that respects those constraints must also have states with representational content to count as computational. In their turn, non-semantic theories do not rule out the possibility

[26] Defenders of the semantic view include: Fodor (1975), Peacocke (1994, 1999), Grush (2001), Shagrir (2001, 2006, 2012*a*,*b*), O'Brien & Opie (2009), O'Brien (2011), Ladyman (2009), Sprevak (2010). Egan (2010, 2014*b*) does not appeal to external content, but she belongs to this category inasmuch as she appeals to mathematical contents in order to individuate computations. Churchland et al. (1990) also embrace the view that computation essentially involves representation; however, they claim that concrete computation is observer-relative insofar as it depends on interpreting systems as manipulating representations. Rescorla (2012*b*, 2013, 2014*b*), on his turn, defends a middle ground view: concrete computation, according to him, is sometimes individuated by means of semantic properties, while in other cases it does not essentially involve representation.

[27] Maudlin (1989), Copeland (1996), Dresner (2010), Rescorla (2014*b*).

[28] Klein (2008).

[29] Chalmers (1996, 2011), Chrisley (1995), Scheutz (1999, 2001).

[30] Piccinini (2007*b*, 2008*a*, 2015), Fresco (2014, 2015), Milkowski (2012, 2013).

[31] An exception is, for instance, Ladyman (2009), who conjoins the simple mapping view with a representational requirement.

that computations may be carried out over representations; their claim is merely that even in that case computations are not individuated by their semantic properties — computations do not involve representation essentially.

The debate between non-semantic and semantic accounts of concrete computation revolves mainly around issues such as the following:

- which family of theories, if any, does justice to how computational explanation works in computer and cognitive science?
- which family of theories, if any, provides a non-trivial account of concrete computation?
- which family of theories is able suitably to capture the domain of computational systems, including the right systems and excluding the wrong ones[32]?

To illustrate the debate, I will briefly examine one of the most influential non-semantic theories of concrete computation: Chalmers' causal topology view. I will then pit it against four core families of arguments for the semantic view. I will argue that three of them fail, while one does put non-semantic views in dire straits. Assessment of how fatal this is for non-semantic theories will have to wait for the next chapter, in which I will present and defend my version of a recent type of non-semantic theory, the mechanistic view.

3.3.1 Chalmers' causal mapping theory

In a series of papers, David Chalmers (1995, 1996, 2011, 2012) has defended a sophisticated causal mapping account of concrete computation. Causal mapping theories complement the simple mapping view with constraints on the causal structure of the implementing physical system.

A natural rejoinder to the Putnam-Searle trivialisation objections is to point out that the state-transitions taken into consideration do not support counterfactuals[33]. Putnam's triviality arguments employ only material conditionals in describing state-transitions, which have no modal force. However, a theory of concrete computation is concerned not only with occurring state-transitions, but also with non-occurring ones. The state-transitions must support counterfactuals of the form 'if T were to enter physical state a then it would transition to physical state b'[34]. The proof presented in section §3.1 does not go through if we have these stronger state-transition conditionals in mind. For the arbitrary system to which the computational description is mapped is open, and vulnerable to all sorts of external interference. Any interference, even of an atomic particle, would change the overall physical state of the system and its behaviour. Putnam's proposed mapping is silent on these alternative trajectories in state-space of the physical system. It only works for the specific circumstances in place during the defined time-interval[35].

[32] Piccinini (2007*b*).
[33] Chalmers (1996).
[34] Copeland (1996), Rescorla (2014*a*).
[35] Chalmers (1996).

Most theories that take route (c) add factors that support counterfactuals, such as dispositions, causality, functional role. Chalmers' view is based on causal organisation. A physical system implements a certain computation when its causal transitions mirror the state-transitions of the formal description, *i.e.* when its causal structure mirrors the formal structure of the computation[36]. In other words, "a computation is simply an abstract specification of causal organisation"[37].

There are worries that at least for what regards inputless FSAs the appeal to causality is not enough to avoid trivialisation arguments[38]. The same worries apply to FSAs with weak inputs and outputs, while FSAs with strong inputs and outputs are arguably not liable to triviality arguments[39]. At any rate, Chalmers (1996, 2011) argues that the FSA is not adequate as a computational formalism for cognitive science. As we have seen, FSAs have monadic states — there is no structure to their internal states. The formalism is implausible as a computational description of computers and cognitive systems, systems that have considerable internal complexity relevant to their computational capacities.

Chalmers suggests that Combinatorial-State Automata, or CSAs, are the appropriate computational formalism for computer and cognitive science[40]. Contrary to FSAs, at any point in time CSAs are in a structured combinatorial state described by a vector $[S_1, \ldots, S_n]$. Each S_i is a separate component of the automaton, and it can be in one of a finite number of states. The states of the components are thus substates of the whole automaton. State-transitions are described as going from vectors of inputs and internal states to vectors of internal states and outputs, $[I_1, \ldots, I_k], [S_1, \ldots, S_k] \rightarrow [S'_1, \ldots, S'_k], [O_1, \ldots, O_l]$. CSAs have complex internal structure, and are thereby in a better position accurately to describe complex systems such as computers and brains.

Briefly, for a physical system to implement a CSA there must be a decomposition of it into independent elements[41] in substates whose causally-mediated state-transitions map onto the substates and state-transitions of the elements determined by the CSA description[42]. The recourse to internal structure makes the requirements on the implementing physical system much more demanding than in the case of FSAs. Causal, counterfactual-supporting state-transitions must be appropriate for all the substates of the physical system so that they can count as implementing a CSA. It does not suffice that a system be sufficiently complex in order to implement a computation, as per the

[36] Chalmers (1995, 2011).

[37] Chalmers (1995, p. 396), Chalmers (2011, p. 331.).

[38] See Chalmers (1996), Sprevak (2012).

[39] Chalmers (2012, pp. 236-7.)

[40] In his (2012), Chalmers endorses Milkowski's (2011) suggestion of replacing CSAs with Abstract-State Machines, or ASMs. Nothing of moment for our limited purposes follows from this, so I will ignore this recent change of mind.

[41] As Sprevak (2012, p. 138) points out, this requirement plays an important role in avoiding trivialisation arguments insofar as it keeps CSAs from being reducible to FSAs. Chalmers cashes it out as a matter of spatial distinctness, which is admittedly problematic, given that it rules out from the account computational systems that use spatially overlapping components to perform computations. The account lacks a satisfying way of determining what counts as an independent element in the physical system, as we will see below.

[42] For a formal characterisation see Chalmers (1995, p. 394), Chalmers (1996, p. 325), Chalmers (2011, p. 329.)

simple mapping account. The implementing system must have the right complexity. The causal goings-on inside the system must mirror in a fine-grained way the detailed computational description of a CSA. The causal mapping view based on the CSA formalism puts triviality worries to rest: few systems will have the required fine-grained causal structure required for implementing a CSA.

However, even though concrete computation is in most interesting cases not trivial, pancomputationalism still ensues, albeit of a limited sort. On this view, every physical system implements at least one computation, described by a single state FSA. As the causal complexity of a system increases, it implements more and more complex computations, as well as the simpler ones. But it is not true, as it is for the simple mapping account, that every physical system implements every computation. Rather, complex computations described by CSAs are implemented by few systems. Therefore, computational implementation is not a trivial matter, and appeal to computation is not vacuous — while claiming that a system implements a simple FSA may be trivial, to say that a physical system implements a complex enough computation, on the other hand, is to make a substantial claim.

Chalmers' causal mapping view of concrete computation avoids unlimited pancomputationalism and embraces in its stead limited pancomputationalism[43]. While it is not true that every system implements every computation, as the Putnam-Searle triviality arguments would want, every physical system performs at least one computation. However, whether a physical system implements a complex computation is far from trivial.

In this sophisticated causal mapping theory, computational descriptions capture the abstract causal organisation of a system, what Chalmers (2011) calls its 'causal topology'. It is abstract insofar as it leaves behind the physical nature of the elements in the system, focusing exclusively on their causal relations. Two systems made out of radically different components, such as empty beer tins and silicon chips, can have the same abstract causal structure — the same causal topology — as long as the structures of causal relations between their elements are the same. Computational descriptions capture this similarity: two such systems implement the same computation. Computational explanation is causal explanation that abstracts away from the physical details of the system under investigation. In some cases, such as that of computers and perhaps cognitive systems, this is a relevant, if not the most relevant, type of explanation.

Chalmers' view is promising in that it significantly improves on the simple mapping view of computation in ways that do not endanger its naturalistic status. Nonetheless, it has been the target of various serious criticisms[44]. I will only examine some of the shortcomings of the view, the ones that strike me as most cogent.

First, as for the simple mapping account, no constraints are placed on how to select and group the states of the physical system to be mapped onto the computational description. The only requirement is that the substates of the physical system be values

[43] As does Scheutz's (1999) causal mapping account.

[44] See articles in the Journal of Cognitive Science, from vol. 12, issue 4, to vol. 13, issue 3, for comments and Chalmers' reply. See also Scheutz (2001), Shagrir (2012*b*), Rescorla (2013).

of independent, separate components of the system, understood as spatially distinct[45]. Though Chalmers occasionally flirts with appeals to naturalness in order to block potential trivialisation arguments that may exploit arbitrary groupings of physical states, as he himself admits these appeals are rather vague and obscure. At any rate, he claims that "there is some inevitability in the appeal to naturalness in understanding the notion of implementation"[46]. This move is tantamount to accepting an obscure notion at the basis of the account — a notion, furthermore, that is supposed to play an important role in blocking the triviality objections that were the main motivation for complementing the simple mapping account to start with. Therefore, *desideratum* (i) on theories of implementation, *i.e.* that the notion of concrete computation be clarified, is endangered. If alternative theories of concrete computation are available that do not have recourse to vague appeals to naturalness (or to other similarly obscure notions), they should be preferred to the foregoing causal mapping account.

These considerations also lead to a related worry. No constraint on the relevant causal levels is proposed. A complex physical system will thus perform several computations at the same time — from the more complex to the simpler ones — and will do so for each of its causal levels, from the atomic to the molecular to the macroscopic. According to the foregoing theory, any complex system will perform a significant amount of different computations at the same time. How to choose which one to attribute to the system[47]?

Chalmers accepts that any complex physical system performs multiple computations at the same time, and admits that this may lead to some interest-relativity of computation. A physical system has many different causal structures, "corresponding to different ways of grouping states of the system into state-types"[48], and each of these structures implements different computations. Which causal level to focus on, and which computation performed at that level to attribute to the system will depend on the explanatory interests brought to bear from case to case[49]. This is to some extent analogous to the move recommended by the pragmatist, but with a crucial difference: the pragmatist accepted the triviality of computation, while Chalmers does not. While the pragmatist had to appeal to explanatory interests in order to save attribution of computational states and processes from being vacuous, the appeal to explanatory interests in the foregoing theory is doing much less work. The claim that a physical system performs a (complex) computation (or several) is not trivial as it was in the simple mapping account.

Nevertheless, we still get the counterintuitive result that a complex enough physical system implements a large number of computations — rocks, walls, and pails of water included. Consider once again the rich causal goings-on inside a wall. In the macro-

[45]Which is in itself problematic. See footnote (41).
[46]Chalmers (2012, p. 237.)
[47]Piccinini (2015, pp. 22-3.)
[48]Chalmers (2012, p. 230.)
[49]Scheutz (2001) proposes a somewhat different solution to the problem: he claims that concrete computation is relative to a given physical theory which determines beforehand the appropriate groupings of states. This move, however, when applied to concrete computation in the cognitive sciences introduces worries similar to the ones examined in section §3.2.

scopic level, it will likely implement only simple computations at particular intervals of time, such as one-state FSAs. On the molecular level the implemented computations will be more complex, and arguably even more so at the atomic level. Therefore, on Chalmers' causal mapping view, we should objectively attribute complex computations to relatively simple physical systems. This, as well as the limited pancomputationalism that characterises the view, is at odds with *desideratum* (ii) on theories of concrete computation: the account does not narrow down the domain of computational systems, setting them apart from non-computational systems.

It must be said that this is no decisive reason to reject the causal mapping view. By embracing pancomputationalism, proponents of the theory are indeed repudiating (ii). However, alternative accounts that respect that plausible *desideratum* should be preferred. As Piccinini (2015, p. 55) notes, our sciences generally assume that there is a difference between systems that are computational, and systems that are not — computer science, for instance, is dedicated to the study of very special systems, not rocks, walls, and pails of water. A less revisionary view of computation would therefore be superior to the foregoing, *ceteris paribus*.

The failure of Chalmers' theory of implementation to respect *desideratum* (ii) might have yet another source: the view arguably does not differentiate between causal and computational explanation[50]. As we have seen, computational explanation, on the foregoing, is nothing more than an abstract form of causal explanation, one in which the physical details are left behind, and only the structure of causal relations between elements is brought to the fore. The causal mapping account denies the distinctiveness of computational explanation — computational explanations are available whenever causal explanations are.

This is true only to some extent, however. Even though every system performs computations, only in some cases computational explanations provide appropriate explanations of the behaviour of physical systems. Computational explanations are suitable for explaining the behaviour of physical systems for which the physical details are not relevant. Computational explanations of digestion are not cogent — even though the digestive system performs computations — because to explain digestion we cannot abstract away from the physical nature of the elements instantiating the causal structure[51]. Digestion only takes place when specific types of material participate in the process. A description of its causal topology does not explain how food gets digested. Therefore computational explanation is, at least to some extent, distinct from causal explanation. The latter may be appropriate whenever the former is, but the converse is not true.

The account also has trouble handling *desideratum (iv)*, namely the normativity of computation. With its strong reliance on causal processes, which are not normative, it seems that the view lacks the tools necessary to make sense of the notion of miscomputation. What we would generally consider a malfunctioning computer is, by the causal theorist lights, simply performing a different computation than the one we had expec-

[50] Piccinini (2008*a*, 2015), Ladyman (2009).
[51] Chalmers (2011, p. 332.)

ted. Therefore, it looks like normativity can enter the picture only in a non-naturalistic way, *i.e.* by means of the intentions and expectations of human beings who impose that certain computations be seen as 'correct' — and deviant ones as miscomputations.

In sum, Chalmers' causal mapping account of concrete computation is markedly superior to the simple mapping account. It avoids trivialisation of concrete computation, as well as vacuity of computational ascription, without having recourse to interest-relative considerations, as did the pragmatist. Hence it satisfies *desideratum* (iii), which the pragmatic view endangers. However, the view has problems with the other three *desiderata*. Though it avoids unlimited pancomputationalism, it makes every physical system into a computational system, flouting (ii). Furthermore, the appeal to vague naturalness considerations[52], even though limited, jeopardises (i). Finally, it lacks the tools to account for the normativity of computation in naturalistic terms. Therefore, despite the progress made in comparison to the simple mapping account and the pragmatic view, the refined causal mapping view of concrete computation exhibits several reasons for dissatisfaction. Let us turn, for now, to the opposite side of the fracture and see whether appeal to semantic properties can provide a more satisfying account of concrete computation.

3.3.2 The semantic view of computation

Proponents of the semantic view of concrete computation claim that computation essentially involves representation: to compute is to manipulate states endowed with some form of content. In the dialogue with defenders of non-semantic theories, four main strands of argument in favour of the semantic view can be identified:

Descriptive accuracy the semantic view is more truthful to the practices of computer and cognitive scientists, both historically and currently.

Explanatory adequacy the semantic view offers an account of concrete computation that is more suitable to the purposes of computer and cognitive science than non-semantic views.

No pancomputationalism the semantic view avoids both unlimited and limited pancomputationalism, as it denies that all physical systems are computational. Only systems that represent can be computational.

No multiplicity of computations the semantic view provides the tools to curb multiplicity of simultaneously implemented computations, and ascribes unique computational descriptions to the systems of interest.

I will briefly examine each of these argumentative strands.

[52] There is also an appeal to 'normal background conditions' in order to block trivialisation, which is similarly problematic. See Chalmers (2012, p. 235.)

Arguments from descriptive accuracy

Arguments that hinge on descriptive accuracy considerations are moved by, among others, Peacocke (1994), Sprevak (2010), Rescorla (2012*b*, 2013)[53]. The claim is that computer and cognitive scientists understand computation as essentially involving representation. Therefore the semantic view is superior to the non-semantic one insofar as it is not revisionary about the conceptions and practices of these scientific fields. In the case of computer science, it is often claimed that Turing's original notion of computation is essentially representational[54]. Moreover, register machines and implemented programmes, it is argued, also require positing states endowed with representational content[55]. Analogously, computational explanation in the cognitive sciences is seen as having recourse to representation.

The matter is extremely controversial: proponents of non-semantic views flatly deny that computer and cognitive science work with a notion of concrete computation that essentially involves semantic properties[56]. The dispute is particularly complex to assess, given that non-semantic views deny only that computation necessarily involves representation, while being open to representation playing an important role in many cases. So examples in which computer and cognitive scientists employ or seem to employ representational talk in their endeavours is not enough to tip the balance toward the semantic side, as they are compatible with the claims of both sides.

Furthermore, it is difficult to affirm, when perusing the scientific literature, whether the notion of representation is playing a constitutive role in characterising concrete computation, or is taken to be accidental, or even merely metaphorical, or heuristic — a way for us to understand and keep track of what is going in the computational system. Interpretation of the founding fathers of computational theory as using a semantic *vs.* a non-semantic view of computational individuation is fraught with these same difficulties. Finally, the predominance from the early days of the semantic view as the default framework in the cognitive sciences cannot be taken to be an argument in its favour. Its detractors may reject the case for the better descriptive accuracy of the semantic view by arguing that it was simply the assumed, unquestioned paradigm. Critical scrutiny may show that appeal to contentful states is doing no essential work, at least for what regards individuating computational processes and states.

Therefore, I take the argument from descriptive accuracy to be at best inconclusive.

Arguments from explanatory adequacy

Advocates of the semantic view often invoke the putative superiority of their account over non-semantic theories with respect to its explanatory adequacy both in computer and cognitive science[57]. Sprevak (2010) claims that non-semantic views cannot accom-

[53] In the case of Rescorla, who does not subscribe to neither non-semantic nor semantic views of computation, this line of argument is used to attack non-semantic views.

[54] Peacocke (1994, p. 320), Peacocke (1999, pp. 197-8), Sprevak (2010, p. 268).

[55] Rescorla (2012*b*, 2014*b*).

[56] For instance, see Chalmers (2011, p. 334), Piccinini (2008*a*, pp. 211ff.), Rescorla (2014*b*, p. 1298), Piccinini (2015, pp. 125-6).

[57] See Peacocke (1994, 1999), Sprevak (2010), O'Brien (2011).

modate the possibility of two very different systems performing input-output equivalent computations. Imagine, Sprevak suggests, one system that takes two Roman numerals, and outputs another Roman numeral; and another system that takes two Arabic numerals, and outputs another Arabic numeral[58]. Suppose further that both systems are computing the addition function. Since the two systems are working with different inputs, outputs and, importantly, different mathematical notation, the physical goings-on will substantially differ in their performance of the addition function.

Sprevak claims that the only way to see the two systems as computing the same function is by characterising inputs and outputs in terms of what they represent. There may be nothing else that the inputs and outputs of the two systems have in common, nothing non-semantic that could reveal the two systems as being input-output equivalent, *i.e.* as computing the same function[59]. The semantic view is the only one able to capture this important similarity between the two systems. Therefore, Sprevak claims, it should be preferred over non-semantic views, which make such similarity invisible.

While Sprevak places the weight of the argument on the non-semantic differences between the inputs and outputs of the two systems, and on the exclusive capacity of the semantic view to capture their similarity, I think that the difference in notation is also important. Such a difference motivates the rejoinder from the proponent of non-semantic views. Given the difference in notation, the two systems will have to go through different causal processes in order correctly to compute addition. The algorithms employed over Roman and Arabic numerals will be different, the state-transitions will be distinct. It seems thereby that the two systems are performing different computations, even though they are both computing addition[60]. Furthermore, the non-semanticist argues, it is the finer-grained level that matters for computer science — it is at this level that algorithms and computational processes are precisely characterised. On knowing only that a system computes addition, we still ignore how it does so. And the 'how' question is the one of most interest to computer scientists.

The semanticist can fall back to the view that their account is more appropriate to the explanatory purposes of the cognitive sciences[61]. The argument is nicely formulated by Peacocke (1994, p. 304): "It looks for all the world as if much theorising in psychology attempts to explain particular intentional, content-involving properties of a subject [...] Yet offering computational explanations of these intentional properties would just involve a mistake of principle, if the non-semantic view of computation is correct". The point is intuitive enough: cognitive science is interested in explaining representational capacities of organisms; thereby non-semantic computational explana-

[58] Take the numerals as mere physical shapes, deprived of representational content.

[59] Piccinini (2008*a*, p. 223) dubs this line of argumentation 'argument from the identity of computed functions'.

[60] See Piccinini (2008*a*, pp. 223-225) for another argument against this line of attack. Piccinini argues that while the semantically-individuated computed functions will be the same, at a finer-grained, string-theoretic level the computed functions are different — in the former, the systems are IO-equivalent; in the latter, they are not. According to Piccinini, it is the latter that is of interest to computer scientists, and, moreover, it must be presupposed in order to allow the former, semantic understanding.

[61] See Peacocke (1994, 1999). Piccinini (2008*a*) calls this line of defense 'argument from the identity of mental states'.

tion, with its silence on representational properties, seems ill-suited to that aim. The acceptance of externalism about content, a rather common view, provides an additional, related argument: cognitive science is interested in explaining representational capacities of organisms; representational capacities are externally-individuated; non-semantic computational explanation is putatively internalist[62]; therefore the non-semantic view is ill-suited for the purposes of cognitive science.

It is easy to see that the argument is wrong-headed[63]. It assumes that in order to explain a certain property the *explanans* must also feature that property. But this is clearly absurd. The explanation of why certain substances have the property of liquidity under certain conditions makes no reference to liquid molecules, nor the explanation of how certain substances have certain colours under certain conditions makes reference to coloured molecules, or coloured neurons. In the first example, the property is (weakly) emergent, while in the second, it can be seen as both (weakly) emergent and relational[64]. There are plenty of other cases in the sciences in which the *explanans* does not possess the property that is (part of) the *explanandum*. There is no reason to believe that semantic properties are an exception. Indeed, theories that naturalise content try exactly to explain semantic properties in non-semantic terms[65].

Despite the failure of the argument above, a case can still be made that the semantic view of concrete computation meets the explanatory needs of the cognitive sciences better than its competitors. As Sprevak's example of the two addition machines shows, appeal to representation captures commonalities that are invisible to non-semantic views. Therefore, the former allows generalisations that the latter does not[66]. The semantic view can put in the same category a diversity of systems with different architectures, algorithms, and notations by underlining how they all compute the same function. This kind of generalisation seems particularly appropriate for cognitive science, since in many cases fruitful explanations of behaviour abstract away from the specific algorithms implemented, as well as the causal and functional properties of cognitive systems. Ants and desert rats are both capable of path integration, even though they might not implement the same fine-grained computations. The semantic view captures the similarity of the two organisms — they both compute the direct path from a certain location back home — while non-semantic views miss it. This sort of similarity is especially interesting for (some branches) of the cognitive sciences. Consequently, we should embrace a view of concrete computation sensitive to it.

I think that this line of argument from explanatory adequacy is particularly promising. The appeal to generalisations that only semantic individuation of states provides is cogent, insofar as such generalisations are the bread-and-butter of cognitive science.

[62] This premise is particularly problematic, as some non-semantic views of computational explanation have recourse to functional considerations that are not internalistic. See chapter 4.

[63] See Egan (1995, 1999), Piccinini (2008*a*).

[64] Depending, of course, on one's favourite metaphysics of colour.

[65] The counter-argument I am offering in this paragraph is stronger than it need be to secure the rejection of the semanticist argument. It suffices to counter, as Egan and Piccinini point out, that the fact that the *explanandum* is individuated in a certain way (*i.e.*, semantically) does not entail that the *explanans* must be similarly individuated.

[66] See Peacocke (1999), Rescorla (2013).

Nevertheless, it remains to be seen whether these are *computational* generalisations, rather than something else. If, as I believe to be correct, computational explanations bring light not only to what the system is doing, but also to how it is doing it, the generalisations above, though useful, are not of the right fine-grainedness to count as computational[67]. This does not expel such generalisations from cognitive science, but rather makes clear that computational explanation does not exhaust its explanatory tools. Representational explanation can also be fruitful, as it nicely captures some interesting generalisations. Nonetheless, it misses the causal and functional details that are the focus of computational explanation. Both forms of explanation are of central importance to the cognitive sciences, and are not incompatible. At any rate, the 'argument from appropriate generalisations' fails to undermine non-semantic views of concrete computation — it merely points to the fact that computational explanation is one useful form of explanation, but not the only one: representational explanation also has a role to play.

Analogous considerations can be made about the role of representation in computer science. Advocates of the semantic view point out that the non-semantic view would entail that much discourse in computer science — which routinely involves appeal to semantic properties — is either wrong or misleading, thus failing with respect to descriptive accuracy. The non-semanticist, I believe, can accept a role for semantic properties in computer science, but resist the semanticist claim that those properties are involved in the individuation of computations. Representational explanation can be fruitful not only in cognitive science, but also in computer science. However, for the reasons adduced above, it should be kept distinct from computational explanation proper.

Arguments from explanatory adequacy, both for what regards computer science and cognitive science, are hence problematic at best. They fail to provide a good motivation to choose the semantic view of concrete computation over non-semantic views. What is more, when applied to computer science, the attack seems to backfire.

Arguments from pancomputationalism

As we have seen in section §3.1 and section 3.3.1, non-semantic views of concrete computation struggle with pancomputationalism, the claim that everything is a computational system. Simple mapping accounts lead to unlimited pancomputationalism — every system performs every computation — and causal mapping accounts, such as Chalmers', result in limited pancomputationalism — every system performs at least one computation. The former is extremely problematic, though it arguably may not be fatal to computational explanation under a pragmatist approach. The latter is also problematic, insofar as it makes everything into a computational system. Furthermore, causal mapping views appeal to obscure notions such as 'naturalness' and 'normal conditions' to keep at bay the risk of being themselves trivialised.

The semantic view seems to have a quick and easy solution to the non-semanti-

[67] See Piccinini (2008*a*), Chalmers (2012).

cist's predicament. Only systems endowed with semantic properties compute, whereby pancomputationalism is avoided[68]. Rocks, walls, pails of water do not represent and are not, therefore, candidates for being computational systems. The semantic view of concrete computation would thus triumph where the non-semanticists foundered: it is able to drive a wedge between computational and non-computational systems, vindicating *desideratum* (ii).

This is of course not only quick, but too quick. First, we need a notion of representation that is less liberal than the non-semanticist's notion of computation. If everything represents, then everything — by the semanticist's lights — computes. Pancomputationalism is back, accompanied by panrepresentationalism. The defender of the semantic view, if they want to appeal to the 'no-pancomputationalism' argument, must thus reject overly liberal notions of representation, such as some versions of interpretational semantics, pure indicator theories, or theories that rely on liberal functional considerations [69]. Proponents of the semantic view must rely on stronger, non-liberal theories of representation if they want to avoid pancomputationalism.

The problem of liberality that applies to the main theories of representation in the literature, as we saw in Part I, becomes relevant here. Ramsey (2007) and Morgan (2014) cast doubt on the adequacy of mainstream theories of representation (and their application to cognitive science) insofar as they fail to distinguish representations from mere causal mediators. If representation is equated with the latter, as seems the case in much philosophy and cognitive science, representation, beside losing its explanatory distinctiveness, becomes widespread. Some form of panrepresentationalism follows and consequently, the semantic view of concrete computation falls prey to pancomputationalism as well.

The problems with the notion of representation point to a general shortcoming of the semantic view. For it appeals to a as yet poorly understood notion in order to ground an account of concrete computation. As such, it introduces considerable obscurity at the heart of the theory, flouting *desideratum* (i). Moreover, it makes the success of the view depend on the success of a theory of representation. Finally, the semantic view closes off, on pain of circularity, one potential route for accounting for representation, namely, grounding it in the notion of computation[70].

Grush (2001) recognises the problems with the appeal to representation in explaining concrete computation, and claims that some mainstream theories of content available, *i.e.* informational semantics and teleosemantics, are 'unworkable'[71]. There may be, of course, other candidates[72]. At any rate, anything resembling a consensus on an adequate naturalistic theory of representation for the cognitive sciences is as yet, after

[68] See Peacocke (1994), Grush (2001), Ladyman (2009), Rescorla (2013).

[69] See Ramsey (2007).

[70] As we will see in Part III, this limitation is particularly troublesome insofar as there are good reasons to hold that this is a particularly promising approach to understanding representation.

[71] See Piccinini (2015, sec. 3.2) for an argument that four influential theories of representation, *i.e.* conceptual role semantics, interpretational semantics, informational semantics, and teleosemantics, are all unsuitable for underpinning a semantic view of concrete computation. He argues that they all must presuppose non-semantic criteria for computational individuation on pain of crippling shortcomings.

[72] Such as Grush's own theory of representation, developed in his (2004).

several decades of philosophical efforts, not in the horizon.

These general considerations are not decisive reasons to abandon the semantic view, though. Rescorla (2013, p. 693) argues that we do not need to wait for a successful naturalistic reduction of semantic properties in order to use them in informing our scientific theories and practices — the explanatory fruitfulness of appealing to those properties suffices to justify their deployment in scientific theorising. Even though I agree with Rescorla on this point, it remains true that the obscurity of semantic properties does undermine the semantic view of concrete computation, and makes it less desirable when compared to non-semantic accounts, if any, that meet the four *desiderata* above.

In sum, the solution to pancomputationalism proposed by the semantic view is quick and easy only in the surface. A suitably non-liberal theory of representation need be provided. This points to a general shortcoming of the semantic view: using representation to explain computation seems to be a bad move inasmuch as the notion of representation seems to be even more problematic than that of concrete computation. This is not a knock-down argument against semantic views, as it may well turn out that computation essentially involves semantic properties, representation thereby being as important a notion for an account of concrete computation as it is to cognitive science. However, all things equal, non-semantic views would have the upper hand if they manage to appropriately explain computation without appeal to representation.

Arguments from multiplicity of computation

To conclude this brief examination of arguments for the semantic view of concrete computation — and against non-semantic views thereof — I will examine the line that I take to be the most promising defense of the semantic view, and the strongest threat to non-semantic views. The idea behind the argument is straighforward. Non-semantic views lead to multiplicity of computation, *i.e.* computational systems perform more than one computation at the same time. However, in both computer and cognitive science, we want to focus on specific computations that are explanatory of the behaviour of systems. Therefore, we need semantic properties to narrow down adequately the computations that systems perform.

Sprevak (2010) offers an illustration of the problem confronting non-semantic views[73]. Consider a logic gate that takes two inputs, and produces one output. The inputs and outputs can be either voltage 0V, or voltage 5V. Suppose that the logic gate outputs 5V if and only if the two inputs are 5V, and 0V otherwise. As Sprevak points out, there is no way to decide whether the logic gate is computing the logical function AND or the logical function OR. It all depends on which voltage is taken to mean '1' or 'true', and which voltage is taken to mean '0' or 'false'. If 5V = 1, we have an AND-gate, if 5V = 0, we have an OR-gate. It seems to follow that unless semantic properties are brought to bear — *i.e.* what each voltage represents — it is impossible to distinguish AND- from OR-gates. The distinction is crucial to computer science, as complex computational

[73] See also Ladyman (2009).

devices are built from large quantities of logic gates wired together in specific ways so as to compute complex logical functions. Any account of concrete computation that fails to do justice to these 'basic distinctions' is hence in trouble.

Shagrir (1999, 2001, 2012*a*,*b*) has put forward a related argument[74]. By taking physical systems implementing logic gates to be tri-stable rather than bi-stable we also reach the conclusion that the same system implements different logic gates. Computational systems are built in such a way as to respond to voltage intervals, rather than precise voltages. This is motivated by the noise intrinsic to such systems, making the attainment of precise voltages difficult and error-prone. Consider again the physical system presented in the above paragraph, with the difference that its inputs and outputs are values inside certain voltage intervals. Shagrir points out that voltage intervals can be differently grouped, making the system tri-stable. Suppose that the system outputs a voltage in the interval 5V-10V if and only if both its inputs are voltages in the interval 5V-10V. In addition, it outputs voltages in the interval 0V-2.5V iff both its inputs are voltages in the interval 0V-2.5V; and it outputs voltages in the interval 2.5V-5V otherwise. If we take voltages in the interval 2.5V-10V to represent '1' or 'true', and voltages in the interval 0V-2.5V to represent '0' or 'false', the system is implementing an OR-gate. However, a different grouping of voltage intervals (0V-5V/5V-10V) makes so that the system implements an AND-gate, as in the preceding paragraph. Once again, we have dual logic gates — two logic gates implemented in the same physical system. With Shagrir's method, other dual logic gates can be built, such as XOR/NAND-gates, and even two different AND-gates[75]!

There appears to be no way of deciding which logical function a logic gate is computing by means of purely non-semantic considerations. It seems that the only way to rule out dual logic gates is by means of their semantic properties. When certain voltages represent certain values, such as 'true' or 'false', the ambiguity is solved. Therefore, the argument goes, computational individuation must rely on semantic properties.

The argument from multiplicity of computations represents a crucial challenge to non-semantic theories of concrete computation. In section §4.4, I will provide a solution, which helps shift the debate in favour of non-semantic over semantic theories. This will require first exploring, and amending, a particularly promising non-semantic view of concrete computation — the mechanistic view. The next two chapters will consist of a presentation and defense of the mechanistic view, which will in Part III play a decisive role in grounding the deflated notion of representation that I want to put forward.

3.4 Concluding remarks

In this chapter, I have presented the problem of making sense of concrete computation, and stressed its importance for the foundations of the cognitive sciences. Cognitive science, as well as the engineering branches of computer science, need a satisfying account of what it is for a physical system to perform a computation, what it is to compute

[74] See also Rescorla (2013).
[75] See Shagrir (2012*b*, pp. 141-2).

'in the wild'[76]. In the next chapter, I will present and defend an amended version of a recent non-semantic view of concrete computation, formulated by Gualtiero Piccinini, Marcin Milkowski, and Nir Fresco: the mechanistic account. I will argue that it is more successful than its competitors in satisfying the four *desiderata*. I will moreover argue that it evades arguments from multiplicity of computations. Treatment of the mechanistic view will conclude my analysis of theories of concrete computation, which will later on prove useful when tackling, once again, the problem of cognitive representation.

[76]The expression comes from Fresco (2014).

Chapter 4

The Mechanistic View of Concrete Computation

The neo-mechanist approach to explanation in science has become increasingly popular in the past years, with successful applications in biology, psychology, and neuroscience. In many cases, New Mechanism has proved to be a promising framework for cashing out how explanations in the special sciences work, and what they should look like. In this chapter, I examine one of the applications of the neo-mechanistic framework: the mechanistic view of concrete computation. The attempt to use the tools provided by New Mechanism to account for computation in physical systems has been developed most forcefully by Piccinini (2007*b*, 2015), Milkowski (2013), and Fresco (2014). The fruits to reap should this endeavour be successful are very significant: extending a fruitful approach to scientific theorising to a domain, computation, that has so far resisted satisfactory naturalisation, and remains problematic when appealed to in scientific explanations. The mechanistic view of concrete computation may provide the much sought-after satisfactory theory of computational individuation and implementation, succeeding where the accounts examined in the previous chapter failed.

The abstract nature of computational explanation introduces a tension in the neo-mechanistic framework, as Haimovici (2013) has pointed out. For one of the defining characteristics of New Mechanism is its insistence on the importance of providing some degree of structural detail about the mechanisms that contribute to explaining phenomena. This requirement seems to be at odds with the abstractness typical of computational explanation. Hence computational mechanists find themselves in a dilemma: either computational explanation is essentially incomplete, or by enriching it with structural detail, we lose its peculiarity, and in particular its medium-independence and multiple realisability.

After presenting the neo-mechanistic framework, and going into detail on its account of concrete computation, I will try and dispel that apparent tension. This involves amending how the mechanistic view of concrete computation is conceived in the existing literature, especially for what regards the role played by the appeal to mechanism. I will argue that New Mechanism makes an essential contribution to our understanding of concrete computation, though for reasons other than the ones normally adduced by

computational mechanists. My approach to the mechanistic view of concrete computation has beneficial consequences for other central debates in philosophy of computation. The argument from multiplicity of computations, I show, can be dealt with satisfactorily within a non-semantic framework.

Suitably amended, the mechanistic view of computation, I will argue in Part III, can provide the basis for a deflated notion of representation. Before going back to representation, however, I will need to delve into the issue of teleological function, a central feature in the mechanistic account, on which its cogency hinges. Teleological functions will be the topic of the next chapter.

Here is how I proceed in this chapter. In section §4.1, I briefly introduce the neo-mechanist approach to scientific explanation, opting for one of its most general formulations. I then present the mechanistic view of concrete computation, in particular as developed by Gualtiero Piccinini, in section §4.2, also bringing to the fore the apparent tension that I aim to dissolve. Next, in section §4.3, I examine Haimovici's objection to the view, which invites a ready reply by the computational mechanist. The reply is unsuccessful against a related worry, and two strategies to answer the challenge are available. The first accepts the terms of the debate set by Haimovici, rendering the recourse to mechanism trivial, and causing the mechanistic view to collapse onto a purely functional view. The second strategy rejects the terms of the debate, embraces the view that computational individuation is functional, but keeps an important role for mechanism in an account of concrete computation — thereby justifying the label 'mechanistic view'. Finally, section §4.4 tackles Dewhurst's (2016) theory of computational individuation, and explores some positive consequences my approach has to issues regarding computational equivalence. In particular, I argue that in accepting my proposal about how to understand the mechanistic view of concrete computation, one of the most powerful arguments against non-semantic theories — the argument from the multiplicity of computations — is put to rest.

4.1 Mechanistic explanation

The neo-mechanicist approach to explanation in science has become increasingly popular in the 16 years since the publication of what may be seen as the manifesto of New Mechanism (Machamer et al. 2000). Though elements of the framework were already being discussed by the likes of William Bechtel and Stuart Glennan, and hints of it can be found even earlier, *e.g.* in work by Jerry Fodor and Robert Cummins, it was with the publication of the 2000 paper by Machamer, Darden, and Craver — often abbreviated in 'MDC' — that the neo-mechanistic approach burgeoned.

A substantial part of the motivation for this movement in the philosophy of science stems from the greater attention paid by philosophers to the workings of the special sciences, such as biology, and the cognitive sciences. The theories of scientific explanation available, such as Carl Hempel's and Wesley Salmon's, tailored as they were to physics, are found wanting when applied to the special sciences. Different types of questions, different answers, and different explanations are sought in the latter, setting them

apart from physics, its explanatory methods and needs. Rather than looking for laws of nature, which arguably do not exist in their fields, biologists and cognitive scientists are interested in uncovering the mechanisms that produce and/or sustain a certain phenomenon or ability. Explanation in the special sciences, in this picture, proceeds by breaking up the phenomenon to be explained into its component parts, what they do, and how they are organised, *i.e.* by unveiling the underlying mechanism that produces, and sustains the phenomenon.

The notion of mechanism, in philosophy as much as in the sciences, is used in importantly different ways, with different theoretical targets in view[1]. My focus is on New Mechanism as a framework for understanding explanation in the (special) sciences[2]. Other uses of the notion, *e.g.* as the basis for an account of causation[3], do not concern me here. One can be a proponent of mechanistic explanation while holding a non-mechanistic theory of causation, such as manipulability theory[4].

Within the domain of theories of scientific explanation, the notion of mechanism has been differently cashed out by different theorists. These different ways of understanding mechanism have consequences for what can be covered by a mechanistic account — *e.g.* must mechanisms involve regularities, or can there be one-off mechanisms; must they be stable or may they be ephemeral; must they be systems, or can their parts be loosely connected? The answers to these (and other) questions help determine the scope of the mechanistic account of explanation: for instance, if there cannot be one-off mechanisms, most explanations in History cannot be mechanistic[5]; and if mechanisms must be systems, many explanations in Physics are ruled out[6].

For my purposes the debate on the characteristics of mechanisms are of secondary importance, though some issues will arise in what follows, especially regarding abstraction from details in mechanistic explanation[7]. I endorse an inclusive notion of what mechanisms are, put forward by Illari & Williamson (2012, p. 120). They propose a characterisation of mechanism that ensures general applicability insofar as it is compatible with different more specific takes on the notion:

> A mechanism for a phenomenon consists of entities and activities organised in such a way that they are responsible for the phenomenon.

This general characterisation does not preclude that specific sciences work with a more restricted notion of mechanism. The objective is highlighting what there is in common among sciences when it comes to appeal to mechanistic explanation[8]. As Craver & Tabery (2016) point out, at the bottom of the various proposed ways of understanding mechanism lie the notions of parts (or components), causings (activities, interactions), organisation, and phenomenon. Each of these notions can be characterised in different

[1] See Andersen (2014*a*,*b*), Levy (2013), Moss (2012).
[2] What Levy (2013) calls 'Explanatory Mechanism', and Andersen (2014*a*) calls 'Mechanism$_1$ '.
[3] See Glennan (1996, 2010*b*).
[4] See Woodward (2003, 2002), Craver (2007), Milkowski (2013).
[5] Glennan (2010*a*).
[6] Glennan (2010*b*).
[7] See section §4.3; Levy & Bechtel (2013), Piccinini (2015).
[8] Illari & Williamson (2012, p. 120).

ways[9]. The fundamental idea, which Illari and Williamsom's (2012) formulation nicely captures, is that, given a phenomenon that we are interested in explaining, explanation proceeds by decomposing it into its parts, and seeing what the parts do, and how the overall organisation of parts and their activities underlie, or lead to the production, or maintenance, of the target phenomenon.

The *explanandum* phenomenon helps to individuate the mechanism[10]. Mechanisms are of, or for, a certain phenomenon. Mechanistic explanation starts with a phenomenon to be explained, and hopefully generates the mechanism for it[11]. The decomposition of the mechanism into its components, activities and organisation proceeds with that in mind. What components must there be, so that the target phenomenon takes place? Which activities must those components engage in? How do their contributions come together in bringing about the phenomenon? These are the questions that scientists attempting to give a mechanistic explanation try to answer.

Components' activities have functions inside a mechanism insofar as they make a contribution to the overall behaviour of the mechanism. Mechanisms may themselves be functional — they might have functions to perform in the context of an organism or artefact. I will refer to this notion of function, *i.e.* in terms of the causal roles of a component inside a system — as systemic functions (Cummins 1975). Causal roles depend on the activities that entities, or components, perform.

Functional considerations play an important role in mechanistic explanation. In explaining the overall capacity of a mechanism, its decomposition proceeds by identifying the components of the mechanism, as well as their systemic functions that help to bring about the overall behaviour. Structural properties are also relevant in mechanistic explanation. Roughly, while structural considerations deal with the components of a mechanism, and their physical properties (such as size, shape, etc.), functional considerations deal with the activities components perform, their causal powers and how they contribute to the capacity of the whole mechanism (Piccinini & Craver 2011). How to understand functional mechanisms is a matter of ongoing debate, and I will come back to this issue in following sections, and in the next chapter[12].

Some constraints on what count as parts or components must be put in place, on pain of providing an account that is too liberal, in which almost anything would count as a mechanism. Glennan (1996) proposes criteria of robustness and independence: parts must be such that they can be extracted from the mechanism they help to compose and be examined individually, without thereby losing their properties (except perhaps for the functional properties they have inside the mechanism). Woodward (2002, S374-375) proposes somewhat similar criteria by having recourse to a notion of modularity defined in terms of independent interventions — the generalisation that describes the behaviour of each component of a mechanism must be invariant under interventions, and can be intervened upon independently of other components. I will not dwell on this issue,

[9]For a detailed overview, see Craver & Tabery (2016).

[10]Glennan (1996, p. 52), Illari & Williamson (2012, pp. 123-4.)

[11]Or at least a possible mechanism. For the distinction between how-possibly and how-actually explanations, see Craver (2006).

[12]See Machamer et al. (2000), Craver (2001, 2013), Garson (2011, 2013), Piccinini (2015).

though it must be kept in mind that the individuation of the parts of a mechanism has both top-down constraints — the functional decomposition of the mechanism whose behaviour is to be explained — and bottom-up constraints — the parts must be, in some way, entities that exist robustly and independently of their participation in the mechanism.

In sum, mechanistic explanation proceeds by individuating the underlying components and activities, as well as their organisation, that form the mechanism, and unveiling how they bring about the phenomenon to be explained. Many cases involve nested mechanisms — the components of mechanisms are themselves mechanisms that can be decomposed into components, which might on their turn also be decomposable mechanisms, and so on, until a level is reached in which components cannot be mechanistically decomposed. This leads to the multi-level nature of mechanisms, and mechanistic explanation.

For my limited purposes, a more detailed characterisation of New Mechanism as a general framework for scientific explanation is not necessary. My focus is on one of its offshoots, the mechanistic view of concrete computation. According to Milkowski (2013), Fresco (2014), Piccinini (2015), computational explanation is a particularly abstract form of mechanistic explanation, suitable for mechanisms that perform computations. I will be mostly concerned with Piccinini's view of mechanistic computation. His theory, I believe, is the most well-argued account currently on offer, and provides the best candidate for a theory of concrete computation that can supplant its semantic and non-semantic rivals.

4.2 Computational mechanisms

According to Piccinini, computational mechanisms are a type of teleofunctional mechanism. Teleofunctional mechanisms are mechanisms that have teleological functions, that is to say, they have purposes or ends (Wimsatt 1972). The purpose of an engine is to provide power, and purposes of organisms include survival and reproduction. The notion of teleological function is not to be confused with the notion of systemic function. In teleofunctional mechanisms, both kinds of function are relevant. The mechanism has one or more teleological functions, and its mechanistic decomposition, in light of those teleological functions that help characterise the mechanism's capacities, partially depend on the causal roles, the systemic functions, of its components.

The appeal to teleological function makes it the case that teleofunctional mechanisms can go wrong: they may fail to perform their teleological functions due to breakage, inappropriate circumstances, etc. Many mechanisms are not teleofunctional. Though they have components that perform activities that explain a phenomenon, they have no end or purpose — think about planetary systems, the formation of valleys, the water cycle. These systems can be broken down into their components and what they do in order nicely to explain how they work, and why they behave the way they do. Even though their components have systemic functions, the overall mechanisms have no teleological function, and therefore cannot succeed or fail in any substantial way. Planetary

systems can be altered in their workings by the intrusion of a new wandering planet, or by the effects of a supernova. Nonetheless, they do not therefore fail in performing any teleological function. They are mechanisms, but not teleofunctional ones. I will use 'function' to refer to systemic functions, and otherwise I will use 'teleological function'.

Computational mechanisms, according to Piccinini (2015, pp. 119ff.), are a subset of teleofunctional mechanisms. They are those teleofunctional mechanisms that have as one of their teleological functions that of performing concrete computations. Concrete computation, in its turn, is defined as the manipulation of vehicles according to a rule sensitive only to (some of) their physical properties[13]. A rule, finally, is a mapping from inputs (and possibly internal states) to (internal states and) outputs.

This understanding of concrete computation is general enough to encompass digital and non-digital forms of computation — Piccinini (2015) dubs it 'generic computation'. Keeping to such a level of generality is appealing, as it allows various notions of computation to be captured without privileging any one in particular. This is especially welcome in the case of the cognitive sciences, as it is unlikely that the brain, if it is a computational mechanism, computes digitally[14].

Digital computation is a subset of generic computation. To perform a digital computation is to manipulate digits and strings of digits according to rules sensitive only to their physical properties. Digits are medium-independent vehicles characterised by the fact that they can be neatly distinguished by the computational mechanism, insofar as they are discrete, and that there is a finite number of them — an alphabet. Two digits of the same type are processed equally, while two digits of different types are processed differently[15]. In physical terms, digits are realised by equivalence classes of physical states that are treated uniformly by the system. A digit in an electronic computer is an interval of voltage values (*e.g.* 0–5V) to which the system responds in the same way.

There are other subsets of generic computation, such as analogue computation, quantum computation, and, more interestingly, neural computation[16]. However, they are less well understood, so my focus will remain on the more tractable notion of digital computation.

One important property of vehicles in concrete computation is their being medium-independent — most of their physical properties are irrelevant to the computation performed (Haugeland 1985). The rules that govern the changes undergone by the vehicles are sensitive only to some of their dimensions of variation, some of their degrees of freedom. Degrees of freedom abstract away from the physical properties themselves — consisting only of their dimensions of variation — and are characterised in medium-independent fashion. Physical systems made out of completely different materials, from silicon to neurons to vacuum tubes to beer tins, can perform the same computations provided they have physical properties with the appropriate degrees of freedom on which

[13]Piccinini prefers the term 'spatiotemporal properties'. However, it is not clear how voltages, on which modern electronic computers rely, count as spatiotemporal properties. For this reason I prefer the term 'physical'.

[14]Piccinini & Bahar (2013).

[15]Piccinini (2015, pp. 127-8).

[16]See Piccinini & Bahar (2013).

state-transition rules depend. For instance, they may perform digital computation if they can stabilise into two distinguishable classes of physical states to which the rules for vehicle-manipulation are sensitive (and if they are computational mechanisms, *i.e.* mechanisms with the function of performing concrete computations). These equivalence classes of physical states may be voltage intervals, presence/absence of beer tins in a certain location, etc.

Medium-independence is stronger than multiple realisation. As Piccinini (2015, p. 122-3) argues, multiple realisation normally involves further physical constraints beyond having appropriate degrees of freedom. A corkscrew can be multiply realised: it can be made of metal, plastic, steel, as well as different combinations of these (or other materials altogether); and it can have different shapes. A device that is not solid enough, or does not have an appropriate shape, will not be a corkscrew — it will not be able to open wine bottles. Corkscrews are multiply realisable, but they are not medium-independent. Computers, on the other hand, are medium-independent. As long as the vehicles have the appropriate degrees of freedom, their further physical properties are irrelevant to their computational role; for their being manipulated according to rules as defined above[17].

Most functional mechanisms are not computational mechanisms insofar as they involve components that are not medium-independent. Hearts have the teleofunction of pumping blood to the body; and though hearts are multiply realisable — different species have hearts with different physical properties, and there can even be artificial hearts — they are not medium-independent, for reasons analogous to the ones involving the corkscrew.

Vehicles and their activities are arrived at by means of mechanistic decomposition. Given the overall capacity of the mechanism to perform computations, it is decomposed into the entities and activities so organised as to bring about that behaviour. The medium-independence of computational vehicles makes computational explanation a particularly abstract sort of mechanistic explanation. The physical details of the components of the implementing mechanism, be them transistors, valves, or neurons, are not taken into account. What matters is that those components have physical properties with the appropriate degrees of freedom — to which transition rules are sensitive — and be adequately organised.

Computational explanation abstracts away from lower-level, implementational details, sticking to a medium-independent description of the behaviour of the system. In order to provide a computational explanation of a system, its explanatory states must be individuated as computations following one's preferred account of concrete computation — in our case, the mechanistic view. If computational individuation involves the medium-independence of vehicles (among other requirements), computational explanation will have recourse to medium-independent states and processes. Computational explanation is abstract inasmuch as it involves the omission of some detail in its de-

[17] Piccinini & Maley (2014) explore in more detail how the medium-independence of concrete computation leads to the many ways in which computational systems can be multiply realisable.

scription of mechanisms[18].

4.3 The mechanistic view of computation and New Mechanism

The question of abstraction from details is a contentious matter in the literature on the neo-mechanistic approach to scientific explanation. Some neo-mechanists seem to imply that a fully satisfying mechanistic explanation should provide a high-level of detail (ideally complete) on all levels of the mechanistic decomposition of the system[19]. This seems at odds with computational explanation as presented above, as well as with other types of explanation, *e.g.* those that rely on the causal connectivity, rather than the structural details, of mechanisms[20].

Such considerations partially motivate Haimovici's (2013) objection to the mechanistic account of computation. She believes that the computational-mechanist is in a dilemma. On the one hand, if good mechanistic explanations require the most detailed description on all levels of the mechanism, computational explanations are clearly not good mechanistic explanations. By necessarily involving medium-independent vehicles, computational explanations will never be detailed enough to respect that mechanistic norm. If, on the other hand, computational explanations should go into detail at all levels of the mechanism, they would also have to include the implementation details for each computational system, forgoing thereby the medium-independence of computational vehicles. What is distinctive of concrete computation, as well as the scientific value of computational explanations in allowing generalisations not otherwise attainable, would be lost. However, this line of objection is not promising, as Haimovici seems to recognise.

The mechanistic framework need not pose such strict requirements on what counts as a good explanation. These would be at odds not only with actual scientific practice, in which abstraction from details and idealisation play a major role; but would also be detrimental as a strategy for scientific investigation. Not every detail is relevant for explanation, quite on the contrary. An important part of scientific explanation consists in selecting what is explanatorily relevant from what is not for a phenomenon under investigation. Often, including irrelevant detail muddles explanation, and undermines its adequacy — in many cases, 'more is less'[21]. The neo-mechanist approach to scientific explanation would be in dire straits should it require detail at all levels up and down the mechanism. Fortunately, it need not subscribe to such stringent requirements on a

[18] Levy & Bechtel (2013, p. 242.)

[19] Such a position is often ascribed to Machamer et al. (2000) and Craver (2006, 2007), among others, and there is space for seeing Piccinini himself as arguing for it, as some remarks in Piccinini & Craver (2011) seem to suggest. Craver, as well as Piccinini have later denied that they subscribe to this view, and have argued that abstraction from detail and idealisation are vital parts of scientific mechanistic explanation (Craver 2014, pp. 39-40, Piccinini 2015, pp. 124-125). See Levy & Bechtel (2013) and Haimovici (2013) for discussion of the role of abstraction in mechanistic explanation. The confusion may stem from an ambiguity between ontic and non-ontic views of explanation (see Halina forthcoming).

[20] See Levy & Bechtel (2013)

[21] Levy & Bechtel (2013)

good explanation. As Piccinini (2015, pp. 124-5) claims,

> ... mechanistic explanation requires the specification of all relevant structural and functional properties at the *relevant* level(s) of mechanistic organisation — in other words, mechanistic explanation of any given phenomenon requires performing appropriate abstractions from lower level details ... Mechanistic explanation in general requires abstraction, and computational explanation is an especially abstract form of mechanistic explanation — so abstract that computation is medium-independent.

However, there is a related worry that Piccinini's reply does not directly address. Full mechanistic explanation differentiates itself from purely functional explanation by involving not only (systemic) functional considerations, but also detail on the structural properties of the components of a mechanism. Pure functional explanation, by the lights of Piccinini & Craver (2011), is at best a particularly 'incomplete or elliptical' type of mechanistic explanation, one in which 'crucial' detail, and/or most of the structural detail (including the decomposition into physical components) is left out ('mechanism sketches'[22]).

Some structural constraints, according to Piccinini & Craver (2011), remain: the functional analysis of a system places constraints on its organisation and components inasmuch as the latter must be such that they can carry out the functions individuated by the functional analysis. The constraints go in the other direction as well — the structural properties of the system limit which functions it can perform. Since mechanistic explanation, in unveiling the mechanisms responsible for phenomena, makes use of both functional and structural considerations, the two being interdependent, functional analysis is an incomplete form of mechanistic explanation — one in which most structural detail is omitted[23].

These considerations lead to a dilemma in many ways not unlike the previous one. On one hand, if we take computational explanation to be a form of functional analysis, it follows that it is an incomplete kind of explanation — one that should be supplemented with more structural detail. But if computational explanation is supplemented by structural detail, then concrete computation loses its medium-independence. On the other hand, and this is the line taken by Piccinini (2015), computational explanation is to be seen as full-blown mechanistic explanation — one in which all relevant detail at the relevant explanatory level is provided. Given that the relevant explanatory level involves medium-independent vehicles, full computational explanation remains quite abstract — it omits all structural detail, except for the degrees of freedom of the system. This horn of the dilemma is not free from problems.

As computational mechanists admit, the structural constraints posed by computational explanation are extremely weak. Almost all structural detail is left out, and only constraints on the degrees of freedom of the structural components of the mechanism

[22] See Machamer et al. (2000), Craver (2006).

[23] See Shapiro (2016) for discussion and criticism of Piccinini & Craver (2011), especially for what regards their claim that it follows from these considerations that psychological explanation is not autonomous from neuroscience.

are left in place. This weak sort of structural constraint is on a par with the structural constraints posed by functional explanation — there as well most of the structural detail is left out, and only very weak structural constraints are in place.

Consider how functional and structural decomposition would proceed when the mechanism under examination is my laptop computer. The functional decomposition, by referring exclusively to functional components such as 'processor' and 'memory register', abstracts from structural detail. Nonetheless, it does place some structural constraints: whatever plays the role of a processor must be so physically arranged as to do what a processor does, and the same goes for the other sub-capacities individuated by the functional analysis. Analogously, the structural decomposition of my laptop, by mentioning components such as the 2Ghz Intel Core chips, or the 500GB solid-state Flash hard-disk, places constraints on the functional properties of the computer: it limits which functions it can perform in a certain interval of time, and how. Functional and structural properties constrain each other to some extent (Piccinini & Craver 2011).

In computational explanation, the only structural constraint in place, having adequate degrees of freedom, is the one that ensures that the structural components of the computational mechanism can participate in computations, be digits, strings of digits, manipulators of digits, etc. — that is, that ensure that they can play the required functional role. This does justice to the medium-independence of concrete computation. Importantly, this is the kind of weak structural constraint that characterises functional explanation. The abstraction from details that is an essential characteristic of computational explanation is comparable to the abstraction from details found in functional explanations[24].

Moreover, providing any further structural detail beyond the vague ones given by a functional analysis is fatal to the nature of computational explanation. If the computational explanation should mention structurally-individuated components, such as my 2Ghz Intel Core chips, it would immediately foil any attempt at multiple realisability or medium-independence. Computational explanation and functional explanation look therefore much alike. Some fundamental features of computational explanation are identical to features of functional analysis. Those same features are essential to computational explanation, thereby leading to the conclusion that computational explanation is essentially a form of functional explanation.

If, as computational mechanists claim, functional explanation counts as providing mechanism sketches due to the lack of structural detail, then there are good grounds to reserve the same treatment to computational explanations — bringing us back to the first horn of the dilemma. In other words, some fundamental features of computational explanation are identical to features of functional analysis, taken by Piccinini and Craver to be incomplete mechanistic explanations. Those same features are essential to computational explanation as understood by Piccinini's mechanistic view,

[24]See Cummins (1975, p. 764), according to whom in functional explanation, as the functional analysis "absorbs more and more of the explanatory burden, the physical facts underlying the analysing capacities become less and less special to the analysed system [...] this is why it is plausible to suppose that the capacity of a person and of a machine to solve a certain problem might have substantially the same explanation ...". See also Egan (forthcoming).

leading to the conclusion that computational explanation is fundamentally a form of functional explanation, and therefore of incomplete mechanistic explanation. This is at odds with the claim that, given the relevant explanatory level for concrete computation, computational explanations are full-blown mechanistic explanations.

It is of no help to claim that while computational explanation stops at the adequate level of abstraction — for if more structural detail were to be provided, it would cease to be a computational explanation — functional analysis, on the other hand, stops short of providing a more satisfying explanation. For the reasons why functional analysis stops where it does, taking on board mostly functional considerations[25], are analogous to the ones that motivate appeal to computational explanation: capturing generalisations that would otherwise be ignored, and making space for multiple realisability[26].

It is interesting to note that Piccinini, in earlier work, has referred to his view of computation as the 'functional view', even though he stressed its relationships with the neo-mechanistic framework already then[27]. He claims that "... computing mechanisms and their states have functional identity conditions, and ... the functional properties of computing mechanisms are all that is needed to individuate computing mechanisms and their states"[28]. And, in another occasion, he holds that "... computational explanation is a special form of functional (or better, mechanistic) explanation, which applies only to systems with special functional properties"[29].

Truth be told, Piccinini subsequently argues that the term 'mechanistic explanation' is more suitable than 'functional explanation'. However, the reasons he adduces for this move are not particularly deep, especially once one has accepted that functional explanation also poses structural constraints on the realising system. His key motivation for switching to talk of mechanistic, rather than functional, explanation, is that some take the latter to explain only by means of ascribing sub-capacities to the whole system, rather than to its structural components, and their organisation. But, first, this is true only of some types of functional analysis[30]. Boxology, for instance, ascribes capacities to functional components of systems, and not to whole systems. Second, once we accept that functional analyses place structural constraints on the realising system, then they place constraints, albeit weak ones, on structural components and their organisation, as mechanistic explanations do.

In sum, there are 5 claims that when put together in a group (or even in sub-groups) are inconsistent with each other:

1. Good mechanistic explanation tends toward full structural detail at all levels of

[25] I say 'mostly' because I am accepting Piccinini and Craver's (2011) claim that functional analysis does introduce some weak structural constraints.

[26] Functional analyses, though abstract, need not be medium-independent, but are often multiply realisable, as shown in the example in section §4.2. There has been in recent years a rich debate on whether multiple realisability, at least for what regards cognitive states, is true. See Shapiro (2000) and Egan (forthcoming) for a taste of the debate.

[27] Piccinini (2007*a*, 2008*a*).

[28] Piccinini (2008*a*, p. 232.)

[29] Piccinini (2007*a*, p. 107.)

[30] Of the three types of functional analysis that Piccinini & Craver (2011) examine, only one, task analysis, would correspond to the description.

the mechanism.

2. Computational explanation is necessarily abstract, insofar as it ignores most structural detail, caring only about degrees of freedom. (Piccinini 2015)

3. Computational explanation is a type of functional explanation. (Piccinini 2007*a*, 2008*a*)

4. Functional explanations are at best mechanism sketches. (Piccinini & Craver 2011, Piccinini 2015)

5. Computational explanation is good mechanistic explanation. (Piccinini 2015)

As we have seen, Piccinini, as well as Craver, Levy, and Bechtel, reject 1., and for good reason. But 2-5 are still inconsistent. There are two ways to deal with this problem, which I will analyse in turn. The first one, in which I believe Piccinini (2015) falls, accepts that Haimovici (2013) poses a legitimate challenge to the mechanistic view, and tries to reply in a way that is consistent with her assumptions about what the mechanistic view of concrete computation is about. I will show that this strategy fails, and that Haimovici is correct in claiming that she has spotted a problem in the mechanistic account, *given those assumptions*. Some claim or claims among 2-5 must give. I argue that 4. is the one that must be rejected if consistency is to be saved. In so doing, however, it becomes obscure why the mechanistic view is mechanistic at all.

I will argue that these become false dilemmas when the structure of the mechanistic view is properly construed. This construal is in agreement with some of Piccinini's earlier work, and in particular his (2007*a*). But let us proceed in order. Let us first accept the terms of the foregoing debate as they are, let us accept Haimovici's points as a challenge, and see what happens.

4.3.1 Accepting the terms

Apparently the best way to keep the mechanistic view of computation consistent with the overall mechanist approach is to reject the claim that explanations that appeal mostly to functional considerations, such as functional analysis, are only mechanism sketches. It is 4. that seems to be causing all the trouble. Get rid of it, and all is well. If this is done, the mechanist can happily accept that functional analyses and other kinds of highly abstract explanations, such as computational explanation, can also be full-blown mechanistic explanations, provided that they include detail about the explanatorily relevant levels of the mechanism. If the relevant level of explanation is fairly abstract, very few structural details are to be included, *e.g.* constraints on degrees of freedom.

This is consistent with Piccinini (2015, p. 124) when he claims that "... computational explanation counts as full-blown mechanistic explanation, where structural and functional properties are inextricably mixed — because they are mutually constraining — within the mechanism". If we accept that functional analysis also features such a mix of structural and functional properties, as Piccinini & Craver (2011) do, then it follows

that it qualifies as full-blown mechanistic explanation. We should therefore reject the view that functional analysis provides only sketches of mechanism, and on the contrary accept it as a type of full-fledged mechanistic explanation. Otherwise, we fall in the inconsistency highlighted above.

The suggestion is to endorse the following jointly consistent claims: a) Computational explanation is a type of functional explanation insofar as it abstracts away from most structural detail, as much as functional analysis; b) functional explanation is a type of mechanistic explanation — one in which most structural detail is left out; c) computational explanation, and at least some kinds of functional analysis, are full-blown mechanistic explanations inasmuch as they give all the relevant functional and structural detail at the relevant level of the mechanism for the phenomenon to be explained.

What is the price to pay for ditching 4.? It is certainly a concession to functional analysis, and it is doubtful that many proponents of New Mechanism will be willing to be this generous. Part of the motivation for the neo-mechanistic framework is to supplement the shortcomings of functional explanation with more demanding requirements on what counts as good explanation — requirements that involve a certain amount of structural detail. Moreover, those mechanists (if there are any) that subscribe to 1., at least as an ideal target for explanation, clearly cannot endorse the suggestion that functional analysis can be a type of full-blown mechanistic explanation in its own right.

There are reasons to take the concession as not only generous, but also well-motivated. On this picture, functional analysis is a kind of mechanistic explanation — the kind suitable for those *explananda* that, by their nature, involve considerable abstraction from structural details, such as concrete computations, and perhaps some psychological capacities. Even though more structural detail can be provided, thereby unveiling the workings of more levels of specific mechanisms, doing so amounts to losing multiple realisability and medium-independence. It amounts to giving up on computational and psychological explanation, inasmuch as we fail to stop at the relevant explanatory level for those kinds of phenomena.

At this point, one may ask whether the mechanistic view of computation actually has anything to do with New Mechanism. What role does the appeal to mechanism play? Its only role seems to be that of making clear that there is recourse to structural constraints, to the components and organisation of the system, even though they may, as is the case with computational explanation, be rather weak. On this construal, I suggest, New Mechanism does not play a substantial role in the characterisation of computational explanation. It amounts only to the observation that good explanations place some structural constraints on the system sporting the behaviour under investigation — a rather trivial claim[31]. The functional decomposition of a system constrains the structural properties that the system sports: they must be such as to allow the functions described by the functional explanation to be carried out. Analogously, the structural properties of the system constrain what its functional properties are — specific functions can only be performed if the system has appropriate structural components capable of

[31] For a similar argument, see Shapiro (2016).

having those functional properties. As abstract as they may be, functional explanations, the mechanist seems to be urging, cannot float free from considerations about the physical structures that underlie the relevant functional properties. But this is hardly a surprising conclusion, to motivate which we would need the neo-mechanist framework.

In conclusion, if the mechanist about computation buys the terms of Haimovici's dilemma, the appeal to mechanism ends up having a trivial and uninteresting role to play. If one should want it, the appeal to New Mechanism could be safely dropped, leaving the foregoing view of computation unscathed. Concrete computation is individuated by some specific functional properties, as described in section §4.3, and computational explanation is a type of functional explanation in which the only structural constraint on the realising system is that it have the appropriate degrees of freedom. Piccinini's (2015) talk of mechanisms seems to add nothing to the picture other than the rather unsurprising claim that "whether a system implements a given computation still depends on its structural features"[32].

4.3.2 The role of mechanism

Accepting the terms of Haimovici's objection is misguided. The role that the appeal to mechanism plays in the account is different from the one assumed by both participants in this discussion. Once we get back on trail, I argue, the crucial role played by appeal to mechanism in the mechanistic account comes to surface is a way that sets aside Haimovici's (2013) worries.

I suggest that the right way to see the role of mechanism in the mechanistic view of computation is as providing the connexion between abstract computation and world that a theory of computational implementation must deliver. The fact that providing structural detail is part and parcel of mechanistic explanation poses no challenge to the view once it is seen in the correct light. Computation is individuated by functional considerations — it is mainly the capacity to go from inputs to outputs according to rules, which, as we have seen, place structural constraints, albeit rather weak ones, on the realising physical system. Computational systems are physical systems that feature this capacity, or alternatively, that have this function. What the appeal to mechanism gives us is a way of connecting computation and world, providing thereby a theory of concrete computation.

The mechanistic view of concrete computation is best seen as a hypothesis about those systems in the world that actually perform computations — the hypothesis being that such systems are teleofunctional mechanisms. The mechanistic view has it that those physical systems in the world that perform computations, and therefore that can be explained computationally, are tokens of a specific type of teleofunctional mechanism.

Therefore, the amended version of the mechanistic view of concrete computation that I propose has it that computation in physical systems consists in:

1. Manipulation of medium-independent vehicles according to rules sensitive only to their degrees of freedom.

[32] Piccinini (2015, p. 98.)

2. The medium-independent vehicles are components of a teleofunctional mechanism[33].

3. The manipulations that vehicles undergo are activities internal to a teleofunctional mechanism.

4. It is one of the teleological functions of the teleofunctional mechanism to carry out 1.

This is a functional characterisation of concrete computation, despite the appeal to mechanism. It provides very little structural detail, as functional characterisations typically do — it is silent on the physical nature of vehicles and the ways they are manipulated, preserving thereby their medium-independence. However, it makes clear the role that mechanism should play in the account. What makes computational explanation mechanistic is the suggestion that physical computational systems are mechanisms, to which, in consequence, mechanistic explanation applies most suitably. These systems can be mechanistically decomposed in light of their functionally-individuated capacity to perform computations.

The resulting mechanistic view of concrete computation is not incompatible with the constraints posed by competing theories. Rather, it includes mapping as well as causal considerations, into a richer, more constrained, picture[34]. The mechanistic view requires that physical computational systems not only have physical states mappable onto abstract computational states, or that they be causal systems. They must be more than that, they must be mechanisms — organised systems with relatively clear boundaries, decomposable into physical parts that play a role in bringing about the overall behaviour of the system. In addition they must be, at least in Piccinini's (2015) picture, teleofunctional mechanisms, *i.e.* mechanisms that have the teleological function of performing computations.

What the mechanistic view insists on is that computational implementation involves components and activities of mechanisms that lead to and enable the capacity to perform concrete computations. Structural detail can be provided here with no risk, since this is not the dimension in which considerations about medium-independence or multiple realisability are of relevance. Implementations are not medium-independent, they must involve things such as silicon, neurons, valves, beer tins, or what have you. To be a physical computational system is to be a system that can perform computations functionally individuated. The mechanistic view argues that such physical computational systems are teleofunctional mechanisms, rather than mere causal structures, as per the causal account.

This is what Piccinini (2007*a*, p. 108) defends, when he claims, referring to digital computation, that to digitally compute is to manipulate strings of digits according to

[33]There need to be no one-to-one mapping between functionally-individuated vehicles and structural components.

[34]Semantic views, as we have seen, do not provide a means to connect computation abstractly characterised (*nb.* not in the sense of abstraction from details) and concrete computation — they only insist that a semantic constraint be added to the ones offered by non-semantic accounts.

rules. This is a functional characterisation of what it is to compute, one that can be multiply realised. On the same page, he goes on to say that

> of course, there remains the task of explaining how it is that a system is capable of performing a certain computation. This will be done by a mechanistic explanation of the system, which explains how the system performs its computations by pointing at the functions performed by its components under normal conditions and the way the components are organised together.

Mechanistic explanation is what provides the explanatorily relevant functional and structural properties on which computational explanation relies[35]. A system performs a certain concrete computation only if it has structural components that have the degrees of freedom required by that computation. The mechanistic decomposition of a system, involving as it does both functional and structural considerations, reveals their mutual constraints, and helps to pin down the explanatorily relevant structural and functional properties. By considering structural components and their organisation, mechanistic explanation avoids that functional properties be ascribed to arbitrary or hotchpotch groupings of physical states. And by taking into account functional considerations, the contribution of structural components to the overall behaviour of the system is unveiled.

This can be seen in the case of computational devices, such as a personal computer. A functional analysis of the device in terms of its capacity to perform computations leads to its functional decomposition in terms of functional components (*e.g.*, black-boxes with labels such as 'processor', 'memory', etc.)[36]. At the same time, its decomposition into its structural components, their activities, and how they are organised, allows ascribing those functions to one or more structural components of the system, given what they do, and how they influence other components. Moreover, the structural decomposition of the system can help reveal its functions, as well as the functions of the structural components. By examining the structural components of a personal computer, the hard disk, the integrated circuit, the transistors, how they interact and what they do, one may, with non-trivial ingenuity[37], conclude what the overall device's function is, *i.e.* to compute — to manipulate medium-independent vehicles according to a rule.

New Mechanism insists on the interdependence of functional and structural considerations when explaining the phenomenon of interest. We can now see a better reply to the dilemmas presented above than the one offered by Piccinini (2015). Computation is to be explained mechanistically because, when physically realised, it essentially involves teleofunctional mechanisms. This would be question-begging if there were no independent reasons to uphold the neo-mechanistic framework. Fortunately, support for the mechanistic account comes from elsewhere, namely from its meeting the *desiderata* for a theory of concrete computation, as we will see in the next chapter.

[35] See Piccinini (2008*a*, p. 210); Fresco (2014, p. 25); Piccinini (2015, pp. 84-85.)

[36] I will come back to the issue of the ascription of the capacity, or function to perform computations in the next chapter.

[37] See Brown (2014) for a nice cautionary tale.

Piccinini's reply to Haimovici is misguided, as it stresses the wrong aspects to try and solve the dilemma she presents. Computational explanation is not mechanistic because it also places structural constraints, albeit rather weak ones, on physical systems — this is true of many (maybe most) kinds of explanation, even at high levels of abstraction; rather, computational explanation is mechanistic because it pertains to systems that are types of mechanism, and thereby for which the most suitable kind of explanation is mechanistic. The issue rests on whether computational systems are to be seen as (teleofunctional) mechanisms. More on that in the next chapter.

The arguments in Haimovici (2013) and in section §4.3, as well as the explicit reply offered by Piccinini (2015), are beside the point. Computational explanation is mechanistic because, if the mechanistic view of concrete computation is correct, physical systems that compute are mechanisms. This does not in any way impinge on the medium-independence of computational description, or on its functional nature. Mechanistic explanation provides structural detail about computational mechanisms because this is needed to explain how those physical systems are able to compute — it is the way to connect abstract computation and the world, to explain how concrete computation is possible.

4.4 Computational individuation and the multiplicity of computations

The version of the mechanistic view of concrete computation I propose leads to other welcome results. In this final section, I argue that my view helps solve a central issue that has been at the centre of debate in philosophy of computing: the problem of multiplicity of computations, presented in section 3.3.2[38].

Recall that, the argument goes, non-semantic theories of computational individuation, including the mechanistic view, do not have the tools to draw distinctions central to the practices of computer science. Shagrir and Sprevak argue that semantic constraints on computational individuation are needed. The argument from the multiplicity of computations is perhaps the most powerful argument in favour of semantic views of computational individuation against non-semantic views, such as the mechanistic one. Thanks to the points made in this chapter, there is a satisfactory answer to the argument, drawing especially on the considerations brought to bear in the previous section. It is thus time to tackle the argument once again, this time better equipped.

I will focus on Sprevak's version of the argument for expository reasons, as it is considerably simpler than Shagrir's version. Consider an electronic computational device D that takes two voltage values as inputs, and produces one voltage value as output according to the following input-output table (in terms of ranges of voltage values):

[38]The multiplicity of computations problem can be seen as one of the possible arguments leading to a deeper issue, which Fresco et al. (forthcoming) label the 'indeterminacy of computation' problem.

Table 4.1: Device D's input-output table

Input 1	Input 2	Output
0–5V	0–5V	0–5V
0–5V	5–10V	0–5V
5–10V	0–5V	0–5V
5–10V	5–10V	5–10V

The device seems to be working as a paradigmatic logic gate. Logic gates are basic computational building blocks that compute logical functions such as AND, OR, NOR, NAND, etc. At first glance, D seems to be an unequivocal AND-gate. Take voltage range 0–5V to stand for 'False', and voltage range 5–10V for 'True', and we get the truth table of conjunction.

However, as Sprevak points out, if we switch semantic contents, that is, if we take the range 0–5V to stand for 'True', and range 5–10V for 'False', we get an OR-gate — the truth table we end up with is the one for disjunction. Without a decision on what the voltage levels stand for, or represent, so Sprevak argues, there is no way of telling whether D is an AND-gate or an OR-gate[39]. Since logic gates are at the basis not only of theorising in computer science, but also in the engineering of actual computers, semantic properties seem to be required for adequate computational individuation, *contra* theories, such as the mechanistic view, that rely completely on non-semantic properties.

Piccinini (2008*a*) defends the mechanistic view from the argument from multiplicity of computations. He argues that even though D implements more than one computation, a wide understanding of systemic functions, reaching to the immediate context of the computational device (and of the overall mechanism), suffices to determine what the explanatorily relevant computation performed by the device is. Though this answer has some appeal to it, it concedes too much. It concedes that D implements more than one computation, that is, that there is multiplicity of computations — a concession that I believe should not be granted. Moreover, it is not clear that the appeal to the immediate context will be able to curb the multiplicity of computations that Shagrir and Sprevak point out, even when taken in terms of explanatory relevance. Fully dual computational systems, in which all logic gates, as well as the whole system, can be consistently interpreted in two different ways are possible[40]. Though improbable, these systems spell trouble for Piccinini's appeal to wide functional individuation, as in their case, this seems insufficient to eliminate multiplicity of computations — it survives even taking the immediate context of the device, and of the overall mechanism into consideration.

Dewhurst (2016) recently put forward a more promising line of reply. In a nutshell, he accepts that whether devices like D compute a logical function or its dual remains indeterminate by the mechanistic view's lights. However, he claims that this should not worry the computational mechanist, for computational individuation is done at the level of the physical description of the device. The table above is all that is needed to

[39]This is also true of other logic gates, which, due to this property, are dubbed dual gates.

[40]I thank Oron Shagrir and Nir Fresco for bringing this point to my attention.

individuate the computational device, no labelling or ascription of semantic properties is required. AND- and OR-gates are equivalent insofar as computational individuation is concerned. They compute the same function from physical inputs to physical outputs — the patterns of voltage transformation are the same. *Contra* Shagrir, Sprevak, and Piccinini, there is no multiplicity of computations. The indeterminacy lies at a different level, the logical one, that is outside the purview of a theory of computational individuation proper.

This does not mean that individuation in terms of logical functions is uninteresting. On the contrary, it is relevant for many applications in computer science, both in theory and in engineering. But individuation by logical function is over and above computational individuation, and may well rely on wide functions, or semantic properties. Computational individuation is more basic, and non-semantic — it is done at the physical level of the mechanism. Therefore, the charge that Shagrir and Sprevak move against the mechanistic view is misguided. It is true that the mechanistic view does not distinguish AND- from OR-gates (as well as other dual gates), but this distinction is not at the level of computational individuation, for which only the physical patterns of transformation are relevant. Two devices may perform the same computation, but carry out different logical functions depending on contextual and semantic considerations. Computational individuation and logical individuation should be kept distinct. Non-semantic properties suffice for the former, while they might not suffice for the latter.

Admittedly, this picture suffers from a serious shortcoming. It makes computational equivalence impossible, thus also threatening the closely related idea that computations are multiply realisable. As Dewhurst (2016) recognises, "the physical structure of two computing mechanisms is always going to be distinct, and it is unclear whether we can draw any non-arbitrary boundary between the structures that are relevant or irrelevant to computational individuation". It follows that no two computational devices are equivalent, for there will always be physical differences between them that are difficult to rule out as computationally irrelevant in a principled way. But even if we could distinguish the structural properties that are computationally relevant from those that are not, computational equivalence would still be excessively fine-grained, for the physical description of the system is too fine-grained for computational individuation.

To illustrate, take two devices D1 and D2. They work in an analogous way to device D, but with one important difference: for engineering reasons, they have 'cushion' intervals between the voltage ranges relevant for determining the output. Voltages that fall inside these cushion intervals have a 'null' value, and when the device has one such voltage as one of the inputs, it produces a null value, or no output at all. Suppose that D1's cushion interval is 4–5V, while D2's is 5–6V. It follows from Dewhurst's proposal that these two devices are not computationally equivalent, for in the case of D1 the acceptable inputs and outputs are voltages in ranges 0–4V and 5–10V, while in D2's case these are voltages in ranges 0–5V and 6–10V[41]. The two devices have different

[41] I am indebted to Jack Copeland, Nir Fresco, and Oron Shagrir for raising and discussing the points in this and the next paragraph.

physical descriptions, but it seems overly strong to argue that it follows from this that they are not computationally equivalent. Indeed, they have the same number of input and output types (even counting cushion ones), and the former are transformed into the latter by analogous rules of transformation — despite the fact that the processors are sensitive to different voltage ranges.

Similarly, suppose that instead of having different cushion intervals, D1 and D2 were identical if not for being subject to different degrees of noise. Noise makes D1's behaviour unreliable when inputs fall within the range 4.5–5.5V, say, whilst noise interferes with the functioning of D2 when inputs fall within the 4.9–5.1V range. Individuating computation at the physical level would have it that D1 and D2 are not computationally equivalent, despite their striking similarity. In sum, the physical level is too fine-grained to make computational equivalence possible. If we want to save the notion, as we should given its explanatory importance in computer and cognitive science, we need a coarser-grained method for individuating computation (Fresco et al. 2016)[42].

The version of the mechanistic view of concrete computation that I defend in the previous section has the tools to improve over Dewhurst's account in allowing for a meaningful notion of computational equivalence, while keeping to the spirit of his solution to the argument from the multiplicity of computations. In my view, the physical level of description is the wrong one to focus on in order to get adequate, determinate computational individuation — it is too fine-grained to allow for a useful notion of computational equivalence. The physical description gives us the implementational details, but computational individuation takes place at the functional level, in which the only structural considerations at play are having appropriate degrees of freedom. There is a meaningful notion of computational equivalence available at this level of description.

Take again D1 and D2, and their different cushion intervals (or noise levels). While the physical description of the two devices differ, at the functional level their description is identical. The devices respond to two distinct equivalence classes of acceptable physical inputs (voltage ranges), EC1 and EC2, and produce the same equivalence classes of physical outputs (voltage ranges) given the inputs.

The labels are fully arbitrary, and introduced only for ease of exposition. How we label the equivalence classes is irrelevant to computational individuation; what matters is the overall functional profile that defines them. Equivalence classes are defined by input values that lead to uniform behaviour of the whole device — the differences in value to which the device is sensitive, and which are uniformly transformed into new values. For D1, EC1 is the range 0–4V, and EC2 is the range 5–10V, whilst for D2 EC1

[42]Fresco et al. (2016) propose a coarser-grained method for individuating computation that goes some way in the direction I recommend. However, there are fundamental differences between our approaches: they focus on coarse physically-individuated properties, in particular intervals of voltage values grouped into high and low voltages, instead of medium-independent functionally-individuated properties, as in my view; and they fail to draw the crucial distinction between computational and logical equivalence, which leads them to claim that even such coarser-grained individuation methods, as the one I propose, fail to solve the problem of the multiplicity of computations. As I argue, following Dewhurst (2016), once the latter distinction is properly understood, the multiplicity of computations problem becomes considerably more tractable.

is the range 0–5V, and EC2 is the range 6–10V[43]. The input-output tables of D1 and D2, when put in terms of equivalence classes, are identical.

Table 4.2: Input-output table of D1 and D2's functional equivalence classes

Input 1	Input 2	Output
EC1	EC1	EC1
EC1	EC2	EC1
EC2	EC1	EC1
EC2	EC2	EC2

The physical details of the two devices can be glossed over — structural details come in only when we are interested in the particular implementational details of a computational device — as it is the functional description which is of relevance for computational individuation[44]. It follows that D1 and D2 are computationally equivalent: the functional profile from input equivalence classes of physical states to output equivalence classes of physical states is the same.

What the physical states consist in is irrelevant for computational individuation. A hydraulic computational device D3 shares the same functional profile of D1 and D2 if it is sensitive to — and responds uniformly and in the same way to — the same number of equivalence classes of physical states. That those equivalence classes be of ranges of water levels in tanks is interesting when it comes to implementational details, but irrelevant for computational individuation. Thereby, computational equivalence is possible even between systems that work by means of completely different physical principles — and the multiple realisability of computation is preserved.

Indeterminacy of logical function computed still follows. The table above cannot determine whether the devices are AND- or OR-gates (recall that the labels are purely arbitrary, and can be freely switched or changed). Computational individuation, as per Dewhurst's account, leaves logical individuation indeterminate. This is a welcome result, since, as Dewhurst convincingly argues, logical individuation is at least one step above computational individuation. The mechanistic view of concrete computation should not therefore worry about the arguments from multiplicity of computations put forward by Shagrir and Sprevak. What they point out is correct: the mechanistic view does not have the tools to distinguish between dual logic gates. However, such a feat

[43] Alternatively, one could consider there to be three equivalence classes, including the cushion intervals as an equivalence class. This would be the more precise way to go, but I am ignoring this complication for ease of exposition.

[44] In consequence, devices that differ in the number of stable states (e.g. two vs. three), as in Shagrir's (2001) version of the argument from the multiplicity of computations, are never computationally equivalent (Dewhurst 2016). They may be logically equivalent, *i.e.* carry out the same logical function. A bi-stable and a tri-stable device may carry out the same logical function, and thus be logically equivalent, despite not being computationally equivalent given their different functional profiles. Different possible groupings of the devices' stable states, as in Shagrir's argument, are irrelevant to computational individuation: given the different number of equivalence classes of physical states the two devices stabilise on, and are differentially sensitive to, they will always be functionally distinct according to the foregoing account, and therefore not computationally equivalent. This is, I take, as it should be: given their different functional profiles, those two devices differ in their capacity to carry out logical and mathematical functions — having a richer functional structure makes the tri-stable device considerably more versatile.

is not something we should be asking of a theory of computational individuation, for computational individuation takes place below the level of logical functions.

Where I part ways with Dewhurst in on what the appropriate level for computational individuation is. He argues that it is the physical level that allows suitably to distinguish between computational devices. But he has consequently to give up any useful notion of computational equivalence. This is too high a price to pay. In contrast, I argue that computational devices can be appropriately distinguished from each other, or found to be equivalent in an informative way, by focusing on the functional level, in which it is functional, rather than physical, structure, that individuates computational states and processes.

It may be objected that computational equivalence is impossible even when we focus on the functional level, rather than the physical one[45]. It may be argued that the maximal functional profiles of two physical systems will always differ, and thereby that they can never be computationally equivalent. I think that the foregoing account has the means to avoid this objection. For recall that the functional decomposition of a physical system always takes place in light of a target capacity, or teleological function — in our case, the capacity to perform computations. The functional decomposition, and the resulting functional profiles of component computational devices, does not include functional features that are irrelevant to the overall system's capacity to compute. Functional differences between devices which play no role in their general computational capacities are excluded — *e.g.* because they are not relevant to the regimented input-output transformations of equivalence classes of physical states across the system. This makes so that devices that have different physical properties, such as D1, D2, and D3 above, are computationally equivalent, insofar as their computationally-relevant functional profiles are the same[46].

The foregoing proposal hinges on whether there are principled ways of carving the functional structure of a computational device[47]. This is analogous to Dewhurst's (2016) worry about principled ways of carving the computationally relevant physical properties of a system; a worry that, he argues, runs through the whole neo-mechanist framework, and is not a problem specific to its application to concrete computation. Dewhurst suggests, following Piccinini, that only through choices dictated by our explanatory interests can such principles be arrived at. Consequently, a degree of observer-relativity is always in place in mechanistic explanations.

While I agree that there is a crucial worry here, to which a suitable answer must be provided, I believe that a theory of computational explanation has additional tools to deal with it in comparison to other types of explanation tackled by New Mechanism. For the mechanistic view of computation appeals to teleofunctional mechanisms, *i.e.* mechanisms with teleological functions. Teleological functions bestow privileged,

[45] I thank an anonymous referee to *Synthese* and Nir Fresco for bringing this issue to my attention.

[46] Devices that differ in the number of their stable states do not count as computationally equivalent, even though they may be able to carry out the same logical and mathematical functions. This is so because their maximal computationally-relevant functional profiles differ, since the number of equivalence classes of physical states they are sensitive to is different.

[47] I thank Nir Fresco for bringing this point to my attention.

objective capacities on teleofunctional mechanisms, and their components. Hence privileged observer-independent functional and structural decompositions are available for all functional mechanisms, computational ones included. Whether this strategy will bear any fruit depends, though, on whether there are objective teleological functions in the world — a topic for the next chapter.

Chapter 5

Teleofunctional Mechanisms and Concrete Computation

The mechanistic view of concrete computation involves an important appeal to teleological functions. Computational systems are not only functional in the sense that they can be functionally decomposed and explained, they are functional in a further sense: they are endowed with teleological functions. They have, to put it crudely, a purpose, something they are supposed to do[1]. In order to provide a complete mechanistic account of computation, we need to delve into the notion of function, in particular for what regards three crucial questions that arise for the mechanistic approach: what are the grounds to claim that computing systems have teleological functions?; what role does the appeal to teleological function play in the account?; which theories of teleological function are up to the task of filling the role the mechanistic view imposes on the notion of function?

In this chapter, my aim will be to answer the three questions above by making clear why the appeal to teleological function is important for the mechanistic view of concrete computation, which relationships it has with the notion of mechanism itself, and how it makes the account more cogent. Moreover, I will examine several proposals on the nature of teleological functions present in the literature, and assess whether they can play the role that the mechanistic view requires. My aim is not to offer an exhaustive treatment, or to take a position on which account of function is correct. My objective is to offer an existence proof; to show that at least some of the theories of function available in the literature are up to the task set out by the computational mechanist. Thus the cogency of the mechanistic view of concrete computation, with its strong reliance on the notion of teleological function, is preserved.

Before going into depth on the nature and role of teleological functions in concrete computation, I put to rest, in section §5.1, an important worry to the effect that there can be no teleological functions to compute, given the medium-independence of computational states and processes, and the reliance of theories of function on concrete causal powers. That issue circumvented, in section §5.2 I offer an analysis of the roles that

[1] Wimsatt (1972), Neander (1991).

appeal to teleological function plays in the mechanistic view of concrete computation, and what advantages accrue to the view. Teleological function helps the computational mechanist in at least four ways. It helps ensure the objectivity of computation; it helps avoid pancomputationalism; it introduces normativity, making space for miscomputation; and it makes the account fit well with our pretheoretical intuitions about which physical systems are computational. I then move on, in section §5.3, to a closer examination of several theories of function put forward in the literature, divided into two rough categories: dispositional theories, and selected-effects theories. I assess these theories with an eye to whether they are able to complement in the desired way the mechanistic view of concrete computation. I argue that at least two theories of function appear to be suitable candidates: Piccinini's (2015) goal-based theory, and a liberal, broadened selected-effects theories on the lines proposed by Garson (2011). Given a sufficiently robust notion of teleological function, which may already be at hand, the mechanistic view is the best account of how computation takes place in the physical world.

In contrast to the previous chapter, where the systemic notion of function was at the foreground, making me drop the qualification, in this chapter the opposite is the case. Hence I will use 'function' to refer to 'teleological function' throughout, unless otherwise noted.

5.1 Teleological functions and medium-independence

Before we delve into the role of teleological functions in the mechanistic view of concrete computation, a basic worry has to be put to rest. The mechanistic view characterises physical computational systems as mechanisms endowed with the teleological function of performing computations, *i.e.* manipulating medium-independent vehicles according to rules sensitive only to some of their degrees of freedom. There is an internal tension in this formulation. For it is mysterious how a physical system can acquire teleological functions that involve medium-independent vehicles.

Theories of function generally rely on causal contributions of components, present or historical, which explain either their role inside a system (dispositional theories), or their persistence (selected-effects theories). These causal contributions are not medium-independent: they are made by specific components in physical systems, or past instances thereof. It would appear, the worry goes, that theories of function cannot contemplate the bestowing of teleological functions that involve medium-independent vehicles. The causal powers of components and systems that justify ascription of teleological functions involve concrete, medium-dependent vehicles. Even though abstract, medium-independent descriptions of these vehicles and systems are possible, they are not the descriptions that capture how they help bestowing functions on those vehicles and systems. Those descriptions will need to include details about the physical constitution of vehicles, and the systems they help to make up, for it is partly due to such details that a story about how they acquire teleological functions gets off the ground.

The mechanistic view of concrete computation seems in trouble: it claims that computational systems are those mechanisms with the function of performing computations. But the performance of computations, which essentially involves medium-independence, seems to be a bad candidate for being a teleological function of any system — for teleological functions seem to hinge on medium-dependent properties of systems and their components. In this section, I aim at dispelling this worry.

There seems to be an immediate way out for the computational mechanist: to reduce the role that medium-independence plays in the individuation of concrete computation. Computational mechanists insist that vehicles are manipulated according to a subset of their physical properties — those to which the general rule of manipulation is sensitive[2]. It is the causal powers of this subset of physical properties that make the causal contributions relevant for the bestowing of teleological functions — for instance by contributing to the goals of organisms, or being selected for by selection processes. There is nothing problematic in having teleological functions depend on those causal contributions, the computational mechanist could reply.

There are three problems with this line of defence.

First, it seems circular. The subset of physical properties that make the causal contributions of relevance are determined by the general manipulation rule, *i.e.* the computational rule. But such a general rule presupposes that the system has the teleological function of computing that function. As we have seen in section §3.1, there is an indefinite number of ways of mapping causal goings-on in physical systems into computational rules. The general rule that helps select the subset of relevant physical properties of the computational system and its components must be determined by some other factor in order to avoid pancomputationalism. Such a factor, according to the mechanistic view, is precisely the teleological function of the system. It is by having the performance of computations as the teleological function of a system that it is possible to carve it into the physical components and processes that are causally relevant for performing that function. As Piccinini (2015, p. 121) claims, the rule followed by the system is "an abstract (macrocospic) description of the behaviour of a physical computing system when it fulfils its teleological function". Therefore, the computational mechanist cannot appeal to the general rule of manipulation in order to avoid the worry.

Second, if the computational mechanist avoids circularity by refraining to appeal to a general rule, it becomes unclear why the relevant vehicles and processes are computational. The computational mechanist may want to rely directly on a subset of the physical properties of the system, without the intermediation of a general computational rule: the subset that has the causal powers relevant for possessing teleological functions. Once a theory of function gives us the criteria for systems to possess teleological functions, we can search the system for those causal contributions (current or historical) that qualify. Though this is how it works in the general case, it does not work for computational systems.

[2]Piccinini (2015, p. 122.)

For it is not clear why such causal contributions should be seen as involving computational states and processes at all. The function bestowed will be of the same kind as that bestowed on digestive systems, hearts, etc. The latter are not computational systems. Furthermore, by giving up medium-independence, the multiple realisability typical of computations is lost: the causal powers relevant for the possession of the function are not abstractly characterised, they are not primarily functionally individuated, but, on the contrary, include implementational details. As I have argued in section §4.4, focusing on implementational details is misguided when it comes to computational individuation.

Third, the computational mechanist cannot seek refuge in the claim that though implementational details play a role in the bestowing of functions, they count as computational because computational descriptions are possible. Admittedly, computational descriptions of physical systems are always possible[3]. Medium-independence plays a crucial role in driving a wedge between computational and non-computational systems. If it is given up, pancomputationalism looms, and most of the motivation for developing the mechanistic view of concrete computation evaporates.

In sum, this line of reply fails. The problem that got us started is still here: it seems impossible that physical systems can have functions bestowed on them that are to be characterised primarily in a computational fashion. Teleofunctional mechanisms, it appears, cannot have the teleological function of performing concrete computations, since these are charaterised primarily in a medium-independent way, while part of the conditions for function-bestowing, *i.e.* the causal powers of components and systems, are not medium-independent. If that is so, and the medium-independence of vehicles and processes is lost, the computational mechanist has either to embrace pancomputationalism, or to accept that there are no computational teleological functions in the world. As a consequence, the mechanistic view of concrete computation crumbles.

Luckily, all is not lost. There is a way out for the computational mechanistic. Vehicles and processes primarily characterised in medium-independent terms can be involved in the bestowing of teleological functions. Theories of function can deliver medium-independent characterisations of the causal contributions of components and systems as the ones grounding the possession of teleological functions, at least in the special case of computational systems. Let us see how this is so.

It is best to start with the easier case: designed computational systems. In the case of designed systems, it is plausible to think that the intentions of designers play at least some role in fixing their teleological functions. It is plausible to assume that in most cases designers intend to build devices that behave in such a way as to respect abstract computational rules. Typically, the medium-independent specification of the abstract computational architecture of the device to be built predates implementational considerations. Designed computers are then arranged so that their causal behaviour mirrors that computational architecture. In sum, what explains why the physical components of computational artefacts are regimented the way they are are the abstract computational rules to which they are supposed to conform, and for which details of

[3]Piccinini (2015, p. 145.)

the physical constitution of vehicles are irrelevant.

At the basis of the intentions of designers lies the abstract functional characterisation of the computational system. The selection of the best design among the available options is effected on the grounds of to what extent, and how efficiently, a physical system implements that abstract characterisation. Computational physical artefacts are designed and selected in light of the medium-independent functional characterisation that they are built to satisfy. Computationally equivalent computational artefacts are routinely built out of different materials and components, while having the same functional structure. The contributions that computational artefacts make are independent of details about their physical constitution, insofar as functionally equivalent, but differently physically constituted systems, make the same contributions[4]. These considerations lend force to the claim that, as far as computational artefacts are concerned, their teleological functions are primarily characterised in medium-independent terms. Thus the mechanistic view of concrete computation seems able to avoid the objection, at least for what regards designed computers.

This is already quite an achievement. However, computational mechanists also want to account for putative non-designed computational systems, such as cognitive systems. Much of the cognitive sciences, I have insisted throughout, rely on concrete computation as a foundational notion. The mechanistic view would hence better also contemplate the case of (biological) cognitive systems. Things get trickier here, as appeals to intentions of designers are clearly not in the cards. But there is a plausible path available to the computational mechanist. In the case of biological computational systems, there is a sense in which the explanation for why their causal structure is the way it is has to do partly with the computations that it enables. In a way analogous to designed computational systems, computations medium-independently described play a role in determining the teleological functions of those biological systems. There are reasons to believe that the medium-independent characterisations of the causal goings-on in biological computational systems are explanatorily primary in determining their teleological functions.

To understand how this is so, it is useful to make use of the notion of an organisationally invariant property, developed by Chalmers (2011). Organisationally invariant properties are properties that remain invariant as long as the abstract causal structure of a system — what Chalmers calls its causal topology — remains unchanged. In Chalmers' (2011, p. 337) own words, a property P counts as an organisational invariant "if any change to the system that preserves the causal topology preserves P". The causal topology of a system, in turn, is defined as "the pattern of interaction among parts of the system, abstracted away from the make-up of individual parts and from the way the causal connections are implemented". These notions, which are put forward by Chalmers in the context of discussing the role of concrete computation in cognitive

[4]There are of course further issues such as speed of processing, size, etc. Though these may influence to some extent the teleological function of the computational system (*e.g.* in limiting which functions it can compute in a certain interval of time), they do not jeopardise their general function of performing computations.

science, are closely related to the notion of medium-independence that Piccinini (2015) discusses.

What is particularly helpful in Chalmers' discussion is the nexus he builds between organisational invariant properties, concrete computation, and cognition. Concrete computations provide a way of capturing the abstract causal structure of (computational) systems[5]. Crucially, Chalmers claims, mainstream cognitive science accepts that cognitive states and processes involve organisationally invariant properties: it is the special abstract causal structure of some biological systems that makes them into cognitive systems — functionalism about cognition is still the default background position. If that is so, computational description is particularly appropriate to cognitive states and processes, as it captures what is most relevant about them: their causal topology. By the same token, computational explanation is the most appropriate type of explanation for cognitive phenomena[6].

If it is true, as most cognitive science assumes, that cognition is primarily to be explained in terms of causal topology, then the properties that are relevant for bestowal of teleological functions on cognitive systems are also to be primarily characterised in an organisationally invariant way — even though a story that includes the more concrete implementational details will always also be available. Organisationally invariant properties are the realm of computational description. It seems therefore acceptable that some biological systems may have as their teleological function that of going through causal goings-on that are primarily to be described in computational terms — that is, they have the teleological function of performing concrete computations, characterised in medium-independent terms.

Even though the causal contributions that ground teleological functions always involve medium-dependent properties, *i.e.* entities and processes that have (or had) the relevant causal powers due to their physical constitution; in the case of cognitive systems the best way to describe those contributions is in computational terms, abstracting away from details about physical constitution. Thereby, there is a sense in which an explanation of why cognitive systems do what they do — what their function is — is appropriately cashed out in medium-independent, organisationally invariant, terms. More concrete, implementational explanations are always available, and are equally correct. However, they are less adequate insofar as the phenomena under investigation are cognitive abilities themselves — for which explanations are primarily cashed out in organisationally invariant terms.

Another way of putting the point is in terms of counterfactuals. Were cognitive systems composed fully or partly of a different material than neurons and glia, they would function in the same way as actual cognitive systems do[7], provided that the causal topology of the former and the latter were the same (and that the different material

[5] As we have seen in section 3.3.1, Chalmers' view of concrete computation embraces limited pancomputationalism. While for Chalmers concrete computations adequately capture the abstract causal structures of all physical systems, the mechanistic view has it that the domain of physical systems adequately so captured is much narrower.

[6] Chalmers (2011, p. 337.)

[7] Except for, possibly, differences in speed and size.

is suitably connected to, and able to interact appropriately with sensory and motor systems). Thereby, the teleological functions of cognitive systems are not primarily dependent on the physical constitution of the components and processes doing the causal work, but on the causal topology of the systems, which is best described computationally[8].

This line of reasoning does not make the mechanistic view of concrete computation too closely connected to functionalism about cognition, though it does make the former partly depend on the latter, at least for what regards biological computational systems. Functionalism about cognition is only helping account for why, in the case of cognitive systems, it is sensible to believe that their teleological functions can be best characterised in abstract, medium-independent terms. All the other constraints on concrete computation that the mechanistic view puts forward are still in force.

Furthermore, it does not follow that computational explanation and (non-teleological) functional explanation are identical. While, as I have argued in section §4.4, computational explanation is a type of functional explanation, not every functional explanation is computational. For one, computational explanation involves more than functional decomposition *simpliciter* — it requires a very specific kind of functional decomposition in which there are inputs to the system, outputs from the system, and intermediate states that are manipulated according to a general rule. Moreover, many functionally decomposable systems do not qualify as mechanisms, and are therefore excluded from being computational by the lights of the mechanistic view.

Most importantly, the foregoing argument applies only to those systems for which organisationally invariant properties play the primary explanatory role in making sense of their workings. Physical systems of enough complexity, as we have seen in section 5.3.1, can always be functionally analysed in terms of the abstract causal roles of their components. However, in most cases, the functional analysis does not exhaust the explanatory causal contributions of the system, whereas it arguably does in the case of cognitive systems.

Think about digestion[9]. A functional decomposition of the digestive system is no doubt possible. There are subsystems for allowing food to get inside the organism, for breaking food into smaller and more easily processable bits, for transporting food to subsystems dedicated to breaking it further into substances of use to the organism, for absorbing those nutrients, and finally for getting rid of the material that fails to be absorbed. Such a functional decomposition of the digestive system is even quite illuminating: we get a breakdown of the functional components, the work they do, and how that work is subdivided into different stages responsible for different parts of the process of digestion[10].

[8] Chalmers (2011, p. 342.)

[9] Digestion is used as an example by Piccinini (2015) to contrast medium-dependence and medium-independence, and it is used by Chalmers (2011) both to defend limited pancomputationalism from the charge of making appeal to computation explanatorily vacuous, and to contrast organisationally invariant properties with other types of property.

[10] The functional analysis need not stop at this macro-level, but can go on to the causal roles of functional components internal to each subsystem.

However, the properties of those subsystems that are relevant for explaining their causal contributions to digestion are not organisationally invariant properties (or, in other terms, not medium-independent). They may be multiply realisable, but they must be such that they are able to break bits of food, thereby placing important constraints on their physical constitution. Moreover, digestion cannot take place at all without food, which though a variegated category, also places constraints on the physical constitution of its members. Therefore, the causal topology of digestion, its abstract causal structure, though interesting, is not what is primarily explanatory of the capacity of the digestive system to digest. The teleological functions of digestive systems are best characterised in terms of non-organisationally invariant (medium-dependent) properties — properties which are not fully captured by computational description, in contrast with what happens for cognitive systems.

The foregoing considerations lend force to the idea that, when it comes to computational systems, both biological and designed, the factors that ground their having teleological functions are medium-independent, organisationally invariant properties. Such properties are best made sense of in terms of concrete computation.

In sum, the best way to characterise the teleological function of computational systems — designed computers and cognitive systems — is in medium-independent, organisationally invariant terms, rather than in terms of their implementational details. The computational mechanist has thus a plausible rejoinder to the worry presented in this section: physical systems can have as one of their teleological functions that of performing concrete computations.

5.2 *Teleofunctional* mechanisms

Part of the reason why computational systems are to be understood as having teleological functions is that by so doing, a coherent, cogent, and well-motivated theory of concrete computation becomes available. There are four main grounds for having teleological function play such a central role in the mechanistic view: ensuring the objectivity of computation, avoiding pancomputationalism, introducing normativity — thus allowing miscomputation — and doing justice to our pretheoretical intuitions about which systems in the world are computational. I will examine each of them. Any notion of teleological function suitable to the foregoing view will have to be able to fulfil the four requirements below.

5.2.1 Objectivity

Mechanistic explanation involves the decomposition of a system into its components, activities, and organisation. The decomposition takes as its starting point a certain phenomenon to be explained, and often a certain capacity or function that the system is able to perform. For instance, mechanisms are posited to explain the capacity of the kidneys to filter blood. Once we fix the capacity or function to be explained, the search for the explanatory mechanisms behind it can begin, and will be directed and

constrained by the characterisation of the *explanandum*. Choosing the explanatory target, and determining its relevant features, is crucial for mechanistic explanation to take off the ground.

A crucial issue for mechanistic as well as functional explanations is determining the capacities or functions to be explained. Mechanisms are partly individuated by the capacity they aim at explaining, and that motivate the mechanistic and/or functional decomposition that the explanation provides. To some extent, the ascription of capacities and functions to a system depends on scientific interests and explanatory aims. That notwithstanding, the issue remains of whether what is being ascribed to systems, *i.e.* capacities and functions, are observer-independent properties (and therefore interest-based ascriptions may succeed or fail, partially or completely, to capture them), or if they are partly or fully in the eyes of the beholder.

For now, suffice it to note that the mechanistic view of concrete computation advocates an objectivist take on teleological functions. By this means, the computational mechanist is able to defend the objectivity of concrete computations. For a system with one or more objective teleological functions has, as it were, privileged capacities, namely those teleological functions themselves. They are privileged in the sense that they are there regardless of the explanatory interests of anyone. Hearts would have the function of circulating blood even if there were no intelligent life forms in the universe — or so argues the objectivist about functions. The teleological functions of systems are primary targets of mechanistic and functional explanation.

By claiming that computational systems are teleofunctional mechanisms — mechanisms endowed with teleological functions — one of whose teleological functions is that of performing computations, the computational mechanist secures a privileged explanatory target for such systems. Computational systems are to be primarily explained in terms of their capacity to perform computations, since it is one of their teleological functions so to do. They might also be explained with other functions in view, such as that of heating the environment. However, given that heating the environment is arguably not one of the teleological functions of a computational mechanism, such an explanation is of secondary importance — dependent as it is on particular contexts of use — and does not deprive computational systems of their computational nature. Analogously, hearts can be seen as having the function of producing thumping noises in the context of diagnostic procedures carried out by doctors — which is not (arguably) one of their teleological functions. It is the task of a theory of (objective) teleological function adequately to determine what are and what are not the teleological functions of systems, and if they have any.

The mechanistic view of concrete computation, with its appeal to objective teleological functions, makes the computing capacity of computational systems something objective, observer-independent. Such a capacity is a privileged explanatory target, and motivates the functional and mechanistic decomposition of such systems in terms of their performance of computations. This leads to the functional and mechanistic decomposition of computational systems in terms of the components, activities, and their

organisation that give rise to their computational capacities. This decomposition inherits the objectivity of teleological function. It is a non-arbitrary, observer-independent functional and/or mechanistic decomposition[11].

Moreover, the objectivity that the appeal to teleological function bestows on concrete computation puts to rest one of Searle's objections against computationalism in the cognitive sciences. As we have seen in chapter 3, Searle holds that computations are observer-relative — any system can be said to be computational, provided that we find it interesting or useful to regard it that way. Against this view, which risks making claims about physical computing systems vacuous — since any system can be regarded as computational if we should so want — the mechanistic view offers a notion of concrete computation that is objective, non-vacuous, and non-trivial. To attain this result it hinges heavily on the objectivity of teleological functions. Therefore, theories of function able to play the role set by the mechanistic account need respect the objectivity requirement.

5.2.2 Help avoid pancomputationalism

The appeal to function helps with a related problem: pancomputationalism, in its limited, and unlimited forms. For the mechanistic view insists that only mechanisms endowed with specific kinds of functions can be candidates for being computational systems. This is a demanding constraint on which physical systems count as computational. On the mechanistic view, computational systems are mechanisms (thus systems that have to respect some constraints regarding their composition and organisation) which are teleofunctional. In the case of computational mechanisms, one of their teleological functions is that of performing concrete computations. The appeal to function helps curb pancomputationalism in two ways.

First, it provides a robust notion of mechanism. For recall that mechanisms are partly individuated in terms of their capacities. A physical system may undergo different mechanistic decompositions according to the capacities it is taken to have. If functions are not objective, then mechanisms themselves are not objective. Functions, capacities, and mechanisms would thereby depend on the explanatory interests of human beings, not being part of the objective structure of the world. This view, to which I will come back in more detail below, is often dubbed 'perspectivalism'[12]. It has the consequence that potentially any physical system, even of very little complexity, can count as a mechanism, provided that there is an interest in so regarding it[13]. In this watered-down understanding of mechanisms, arbitrary groupings of physical states, as well as agglomerates can be seen as mechanisms of some ascribed, and perhaps outlandish, capacity — provided that they prove in some way to be useful or interesting.

[11] Epistemological limitations may bar us from coming up with an accurate decomposition (even given the relevant explanatory level). Regardless of our epistemic capabilities at any point in time — or of the existence of intelligent subjects — computational systems have a privileged decomposition. See Craver (2014) and Halina (forthcoming) for the distinction between ontic and epistemic notions of explanation.

[12] See Craver (2013).

[13] See Garson (2013).

This outcome brings back the spectre of the Putnam-Searle trivialisation objections: if computational systems are mechanisms, and nearly every physical system can be taken as a mechanism, then nearly every physical system can be taken as computational. Though the risk of trivialisation exists, it does not follow from accepting perspectivalism about mechanisms. Perspectivalism would be compatible with some sort of pragmatism about computation, in which pragmatic constraints would draw the needed distinctions between systems usefully regarded as computing, and those that are not, as per section §3.2. But there remains the lingering worry, fuelled by considerations such as Searle's, that pragmatic constraints will not suffice adequately to curb the liberality that follows from perspectivalism, and will thus lead to the trivialisation of computational claims.

In contrast, a robust, objective notion of function opens the way for a much stronger view of mechanisms. If functions are objective features of reality, and mechanisms are individuated in terms of them, then mechanisms themselves are objective, existing independently of observers and their interests. This does not mean that a physical system will have only one, or even few, functions — and thus contain one or few mechanisms. Physical systems may have different capacities, which will reveal different mechanisms. The point is rather that all those functions and mechanisms will be objective properties of physical systems, independent of observers' aims and interests. Thereby mechanisms themselves are robust, objective features of the world. Importantly, not any physical system will qualify as a mechanism. Mechanisms must answer to special requirements of organisation that exclude from their kin arbitrary hotchpotch groupings of physical states, as well as mere aggregates.

In sum, the appeal to objective functions leads to a non-perspectivalist, objectivist view of mechanisms. If, as per the mechanistic view, computational systems are mechanisms, then arbitrary groupings of physical states are not candidates for computational status. The Putnam-Searle trivialisation objections are thereby at least partially deflected.

There is a second way in which the appeal to objective functions helps the mechanistic view to avoid pancomputationalism. For the claim made by computational mechanists becomes quite stringent. It is not just that computational systems must be mechanisms, but they have to be mechanisms endowed with teleological functions. This rules out systems that, though their components and even the whole mechanisms themselves have systemic functions, lack teleological functions — *e.g.* the water-cycle, and planetary systems.

What is more, not only must computational systems be teleofunctional mechanisms, they must be teleofunctional mechanisms endowed with a specific teleological function: performing computations, *i.e.* transforming input vehicles into output vehicles according to transformation rules sensitive only to specific dimensions of variation of the vehicles (medium-independence). This constraint rules out from being computational all those functional mechanisms that do not have performing computations as their teleological function. Hearts and digestive systems do not qualify. Though they arguably

have teleological functions, none of them is that of performing computations. Hearts and digestive systems do not have the function of manipulating physical states in a medium-independent way[14].

These requirements considerably constrain the domain of mechanisms, let alone physical systems, that can legitimately receive computational explanations. The mechanistic view of concrete computation thereby substantially limits the domain of physical systems that are computational: they must be mechanisms, they must have teleological functions, and one of their functions must be that of manipulating physical vehicles according to rules in a medium-independent way. These constraints disqualify most physical systems from counting as computational, keeping at bay pancomputationalism in its limited, as well as in its unlimited form.

An objective notion of function can provide one or more privileged explanatory targets, *i.e.* the capacities that a mechanism has the function to perform. Moreover, an objective notion of function constrains which mechanisms can be analysed in terms of the performance of which functions. Computational explanation is legitimately explanatory for those mechanisms that have as one of their functions that of performing computations. In their case, describing the phenomenon to be explained as the performance of computations is true to the system's objective nature. In contrast, most systems can receive computational descriptions; they can be seen as performing computations — *e.g.* rocks, pails of water, walls, climatic, as well as digestive systems. Though computational models of such systems can be built — and even be scientifically interesting by, for instance, allowing us to predict their behaviour — as such systems do not have the function to perform computations (and often do not even qualify as mechanisms), they will not count as genuinely computational. The appeal to objective functions helps secure the special, non-trivial nature of computational systems. Though most systems can be seen as performing computations, only some actually do compute: those teleofunctional mechanisms that have as one of their functions that of performing computations.

In summary, the mechanistic view of computation requires an objective notion of function in order to secure the objectivity of concrete computation, and stave off the triviality problems that plague other accounts. Without such a theory of function, the notion of teleofunctional mechanism, and even of mechanism itself, becomes observer-dependent. If mechanisms and teleofunctional mechanisms are not objective features of the world, but rather depend on interests and explanatory practices, it would follow that there is no privileged way of individuating functions, teleofunctional mechanisms, or mechanisms in general. Rocks, pails of water, and walls could be taken to be mechanisms, teleofunctional mechanisms, or even computational mechanisms. At that point, we would be back to the spirit of the Putnam-Searle trivialisation results, in particular in its Searlean version[15]. The success of the mechanistic view of computation, when compared to its rivals, hinges heavily on whether functions are objective, and if they

[14]For more details on medium-independence and multiple realisability, see section §4.2.

[15]Pragmatism about computation in the lines of Schweizer (2014, 2016), and section §3.2 would still be possible.

are, on which theory or family of theories most satisfactorily accounts for the existence of functions in nature.

5.2.3 Normativity

The appeal to function introduces a dimension of normativity in the mechanistic picture of computation. Functions may be performed or may fail to be performed. Moreover, they may be performed at the correct times and at the adequate rates, or they may fail to be so performed[16]. A functional mechanism malfunctions when it fails to perform one or more of its functions, or when it performs at least one of its functions at inappropriate times and at inappropriate rates. This possibility of mistake, of malfunction, makes normative considerations relevant for teleofunctional mechanisms. Kidneys can malfunction if for some reason they are unable to filter blood appropriately, as much as hearts malfunction if they pump blood irregularly.

Similar considerations apply to computational systems. Computational systems can miscompute: they may produce outputs that are at odds with the input-output function it is their teleological function to compute[17]. In analogous ways, pieces of software, such as the LyX word editor, may malfunction when one or more of the computations dictated by the programme are incorrectly carried out — yielding, say, a character that the user did not input. Miscomputation is an important notion in computer science that is often neglected. By giving pride of place to a robust notion of function, the mechanistic view of concrete computation provides the tools to account for the fact that occasionally computations can go wrong, or better put, that there is something it is for a computation to be performed according to the function it is supposed to play, as well as for it to be performed in deviant ways at odds with the teleological functions of the mechanism carrying out the computation.

5.2.4 Intuitive appeal

The appeal to function, as suggested by Piccinini (2015), has some intuitive appeal. There are two main kinds of systems that seem intuitively amenable to computational explanation, and to which the ascription of computational nature seems compelling. Both kinds of systems are normally seen as having teleological functions. The first and less controversial kind is composed by designed computers of all types, of which the programmable electronic digital computer implemented on silicon chips is by far the most pervasive in contemporary societies. Electronic computers have clear, albeit often variegated, functions. They are designed by engineers to perform concrete computations, a capacity tapped on by software developers to make computers compute particular functions instrumental to tasks of interest to human beings.

Electronic computers (as well as other kinds of computational devices) have the general function of performing digital computations for the straightforward reason that they are designed to do so. Computational devices, especially those that work with stored

[16] Garson & Piccinini (2014).
[17] Fresco & Primiero (2013), Piccinini (2015, pp. 148-150.)

programmes, also have more specialised functions — functions to perform particular computations following the instructions implemented in the programmes, and which are on their turn designed to contribute to tasks of interest to the users (or perhaps to other computers in the same network). The designed computers that have become nearly ubiquitous in modern societies are teleofunctional systems, insofar as they are designed and built in order to perform certain functions.

The other kind of systems in the world that are seen as computational, though matters here are considerably more controversial, are cognitive systems of living organisms. Most of cognitive science works on the assumption that cognitive systems operate by performing computations that integrate incoming information, and control appropriate behaviour toward the environment, contributing to the survival and reproduction of organisms. Whether the computational hypothesis will turn out to be correct is still far from settled. At any rate, it has contributed to advances in the young, but quickly developing sciences of the brain and mind. If the idea that cognitive systems are computational should be vindicated, we would have another example of a kind of computational system which is best understood functionally. Talk of function is particularly justified when it comes to organisms and how they are organised. The organs and the processes going on in the bodies and brains of organisms are paradigmatic cases of entities and activities that have teleological functions, most often helping sustain survival and reproduction. Putting together biological functions with the computational hypothesis leads to the view that cognitive systems have the biological teleological function of performing computations.

In sum, the two kinds of systems in the world most likely to be genuinely computational are teleofunctional systems. This observation lends force to the idea that concrete computations are performed by teleofunctional systems, be them biological or artificial. The mechanistic view of concrete computation seems therefore to fit well with the physical computational systems found in the world.

5.3 Theories of function — a (not so) brief *excursus*

A robust and objective notion of teleological function plays a central role in the mechanistic view of concrete computation. It contributes the grounds for claiming that concrete computation is objective, non-trivial, and normative. Moreover, the picture squares well with pretheoretical views on the partial overlap between the kinds of systems in the world that have teleological functions, and those that perform computations. But are teleological functions robust and objective, *i.e.* are teleological functions observer-independent features of the world? Unless the answer to these questions is positive, the mechanistic view of concrete computation is in trouble, for all the advantages above are put in jeopardy.

I do not intend to analyse specific theories of function, if not very cursorily. I will rather focus on circumscribing the families of theories of function that are suitable to fulfil the role the mechanistic view of computation requires. Two main approaches to function have been particularly influential in the debate: dispositional theories, and

selected-effects theories. Though many members of these families fail to respect the *desiderata* set out above, I argue that some versions of the two approaches succeed.

From the 1970's on, seminal works by Wimsatt (1972) and Wright (1973) spurred research on functions, helping to give rise to the vast literature on the subject that developed in the course of the past 40 years. The debate around functions is comprehensibly very complex, and I do not aspire to give anything close to a full treatment. It is though worthwhile to keep in mind at least four dimensions of discussions around functions, revolving around the following questions:

1. Are functions objective features of the world? This is the dispute between objectivism and perspectivalism[18], already briefly touched upon above. According to objectivists, functions are objective features of entities in the world; while the perspectivalist denies that, insisting rather that they are imposed by us in light of explanatory interests.

2. Is function an unified notion? This is the debate between unificationism and pluralism about functions. For the former, a single account can capture what functions are [19], while the pluralist argues that more than one account is needed in order to do justice to the variegated nature of the notion and its usages[20].

3. Which theories of function do justice to scientific employment of the notion? This is a debate about the adequacy of specific theories, or families of theories of function with respect to their ability to capture the use of the notion made by the sciences. For a pluralist, a further question is worth asking: which notions of function are suitable for which scientific enterprises? For our purposes, the question can be stated in a more specific way: which theory, or theories of function are suitable to the cognitive sciences?[21]

4. Which theory or theories of function are best? This question is often asked in debates internal to particular families of theories, in an attempt to develop the most convincing, and less objection-prone version of a theory. But it is also asked in debates between proponents of different approaches, in an attempt to assess which kinds of account are most promising.

For our purposes, question 1. has the most relevance. Nonetheless, all four will surface, to different extents, in the coming discussion. Before going ahead, however, we need at least a crude taxonomy of the families of theories proposed in the rich literature about function. I do not purport that the taxonomy I am offering is the only possible one, or even the most satisfactory. All that matters for my purposes is that, with such a taxonomy in hand, it will be possible more readily to determine which *taxa* correspond to theories of function that the mechanistic view of concrete computation can use.

[18] 'Perspectivalism' is Craver's (2013) term for the view, others, such as Bigelow & Pargetter (1987), prefer to dub it 'eliminativism'.

[19] See for instance Wright (1973).

[20] See Boorse (1976), Godfrey-Smith (1993, 1994).

[21] See for instance Garson (2011).

The driving wedge between the two main families mentioned above — dispositional and selected-effects theories — is the kind of question they take the appeal to function to answer[22]. Dispositional theories propose that functions are invoked when answering questions about how a system, or a type of systems works. These accounts tend to be ahistorical — they focus on the current properties and dispositions of systems, and their components. On the other hand, the selected-effects family of theories takes the appropriate question to be: why is a component of a system there? These theories are historical: they make reference to the past history of systems, and their components in trying to explain how and why they persisted in time, within and across individuals and generations.

To illustrate the distinction, take our running example of the heart. Dispositional theories claim that the function(s) of the heart hinges on how that heart works, what its causal powers are in the context of the system of which it is part. The heart has the function of pumping blood (and perhaps other functions) because that is what it is disposed to do in the context of the circulatory system, and of the whole organism. In contrast, according to selected-effects theories, the proper way of understanding the function of the heart hinges on past instances of hearts, and what they did that made them persist across generations of organisms — thus explaining why a particular heart came to be there. A heart has the function of pumping blood because past instances of hearts in the phylogeny were selected for, and thus persisted because they pumped blood.

This rough distinction in hand, let us proceed to take a brief look at the two families of theories of function in turn.

5.3.1 Dispositional theories

Dispositional (or causal-role) theories of function have recourse to the present causal capacities and dispositions of systems in fixing which functions these and their components have. They offer ahistorical notions of function. Dispositions are understood in terms of counterfactual regularity — were certain conditions to hold, a component (or a system) would manifest a certain behaviour[23]. Hearts have the disposition to pump blood, for when they are introduced in a complex system with specific characteristics, they display blood-pumping behaviour. In order to accommodate stochastic regularities, to be found for instance in the secretion of neurotransmitters by stimulated neurons, the counterfactual regularities that define dispositions must be probabilistic.

Systemic functions play a central role in this family of theories. Roughly, systems and their components have functions insofar as they contribute, or have the disposition to contribute, to a capacity of a larger enclosing system. Their functions are the roles they play, or are disposed to play, in the workings of enclosing systems[24]. Systemic

[22] Cummins (1975), Boorse (1976).

[23] Cummins (1975).

[24] As Cummins (1975, p. 741) puts it: "For something to perform its function is for it to have certain effects on a containing system, which effects contribute to the performance of some activity of, or the maintenance of some condition in, that containing system".

functions, in some theories of this family, are claimed to be the only functions to be had, while other theories claim that they provide the basis for a notion of teleological function.

A decisive aspect that distinguishes various species of dispositional theories is how the enclosing system and its capacities are determined. For a component to have a systemic function, it is necessary to have a way of fixing what the overall system's capacities, to which the component is disposed to contribute, are. I individuate three main paths followed by proponents of dispositional theories in answering this question.

The analytic account

Cummins (1975) puts forward an influential dispositional theory of functions. Functions, according to Cummins, arise when one takes what he dubs the 'analytical strategy' to explanation of dispositional regularities. The analytical strategy starts from a disposition of a system that calls for an explanation, and decomposes it into further dispositions, typically of components of the system. Put together, these dispositions explain the overall disposition of the system target of the explanation. A capacity (disposition) of a system is analysed in terms of the dispositions that give rise to it. These dispositions thereby emerge as functions, insofar as they contribute to bringing about the overall capacity of the system.

Importantly, function-ascription is dependent on an 'analytical context', *i.e.* the overall disposition, or capacity that one sets out to explain. To use Cummins' example, the function of the heart is pumping blood in the context of an analysis of the capacity of the circulatory system to transport nutrients to cells. In a different analytical context, the function of the heart may be different. Hearts have the function of making throbbing noises when the overall capacity of relevance are diagnostic procedures in hospitals, or the capacity of mothers to soothe their babies[25].

Function-ascription is relative to an analytical context. The selection of the analytic context in each case depends on what capacities are singled out as worth investigating. That notwithstanding, functions can be seen as observer-independent features of the world. That physical systems have several different capacities, thereby leading to different functional analyses in terms of the dispositions of their components, is a fact about the world independent of the aims and interests of intentional agents. The latter come in only when it comes to selecting, out of the many capacities — and functional analyses — of a system, the one that is most relevant for the purposes at hand.

It follows from this account that physical systems, be them biological, designed, or else, have several functions. The heart has the function of pumping blood, of making diagnostically useful throbbing noises, of making soothing throbbing noises, and potentially others, since it appears in functional analyses of different overall capacities, playing in each a different causal role — having in each a different function. The case of the heart is not special in this regard. Most physical systems can play different causal roles given the different overall capacities to which they contribute, having therefore,

[25]Cummins (1975, p. 762.)

by Cummins' lights, several functions. Cummins' analytical account makes the notion of function extremely liberal.

Moreover, the account does not provide particularly stringent constraints on which systems are appropriately subject to the analytical strategy[26]. It follows from this approach that any complex enough physical system can be explained by means of the analytical strategy; not only biological or technological systems to which we are pretheoretically happy to ascribe functions[27]. If the water-cycle is the enclosing system, evaporation acquires the function of contributing to the formation of clouds, while clouds have the function of producing rain.

The liberality of the analytic account has two aspects: from one side, systems and components in the world have several functions in so far as they contribute differently to different overall capacities; from the other side, most, if not all, complex systems are good candidates for functional analysis, not only biological and designed systems. This extreme liberality is problematic, if we want to use this notion of function to help ground the mechanistic view of concrete computation. Though the objectivity of functions is preserved, thereby satisfying the first *desideratum* set out above, the liberality that the view entails undermines its ability to help avoid pancomputationalism, and to respect out pretheoretical intuitions about function ascription. Let us briefly look at each of these claims in turn.

The appeal to function plays an important role in constraining the domain of systems in the world that can suitably be seen as computational, the idea being that only systems that are teleofunctional are candidates for being computational. For this move to help avoid pancomputationalism, the notion of function at play must be such as to exclude a considerable number of systems from being functional, and thereby computational. The extreme liberality of the analytical account makes it unsuitable for the task. Given that most, or all complex physical systems can undergo functional analyses, it follows that themselves and/or their components will come out as functional, and therefore as candidates for computational nature. The analytical account can thus contribute very little, if at all, to keeping pancomputationalism at bay. Moreover, the fact that the foregoing account ascribes functions to components and systems that are not biological or designed clashes with our pretheoretical intuitions about which types of systems in the world are functional, thus flouting one of the *desiderata* set out by the mechanistic view.

A traditional objection to the analytical account has it that the notion of malfunctioning becomes problematic, and with it the normativity of functions[28]. A circulatory system that does not work due to damage or malformation still strikes us as having the function of transporting nutrients and cell waste around the body. However, the analytical strategy denies that function to such a system, insofar as it lacks the disposition to perform it. An analysis of a organism with such an ineffective circulatory system

[26]Some criteria to distinguish explanatorily useful from explanatorily useless applications of the analytical strategy are put forth in Cummins (1975, p. 764).

[27]Neander (1991).

[28]For instance, Neander (1991), Buller (1998), Krohs (2009), Garson (2013).

could not ascribe to it the function of transporting nutrients and cell waste, since it lacks that disposition. It has all sorts of other dispositions that can be taken to be functions in different analytical contexts, but those cannot malfunction either. If they were to lack the relevant dispositions, they would lack the ascribed function. Therefore, the analytical account seems to fail the normativity *desideratum* as well.

In sum, the analytical account, though preserving the objectivity of functions, has trouble with the other three core requirements on a theory of function put forward by the mechanistic view of concrete computation. In consequence, it is unsuited to the purposes of the mechanistic view. It is though worthwhile to investigate a recent, and closely related view on the nature of functions, put forward by one of the founding fathers of the neo-mechanistic approach: Carl Craver's perspectivalism. As will become clear, many of the problems that affect the analytical account also affect perspectivalism.

Perspectivalism

Perspectivalism, sometimes dubbed 'eliminativism', has been proposed most prominently by Craver (2001, 2013). Its distinguishing claim is that the capacities of enclosing systems that determine the function of their components are purely a matter of the explanatory interests of human beings. Functions are observer-dependent properties of the world, they are ascribed in light of a perspective taken by human beings in trying to make sense of the causal structure of the world.

What determines what the enclosing system is, and what its capacities are, are choices made in light of explanatory aims. Since enclosing systems and their capacities are the factors that, on dispositional views, determine the functions of the enclosed systems and their components, it follows that functions are relative to the explanatory aims of human beings. As Craver (2013) claims, functions have no place in the ontology of the world, they are conceptual devices used by us in order to delimit and simplify parts of the causal structure of the world that we find, for one reason or another, interesting or useful.

Craver's perspectivalism owes much to Cummins' analytical account of functions. Where it parts ways with the latter is in its insistence that overall capacities, which determine analytical contexts, are mind-dependent properties, whereas in Cummins' view they are observer-independent facts. As a consequence, for Craver functions are interest-relative, rather than observer-independent properties of the world. This account denies functions objectivity. Once transposed into New Mechanism, this view of functions brings with it the consequence, already pointed out in section 5.2.2, that mechanisms themselves are observer-dependent.

A *caveat*: many neo-mechanists accept a sort of perspectivalism for what regards mechanistic explanation[29]. Depending on which phenomenon is chosen as the explanatory target, the same component may appear as playing different roles in its contribution to the overall functioning of the mechanism. If, for instance, one sets off to explain how the circulatory system works in mammals, the heart will be seen as having the function

[29]For instance, Piccinini (2015, p. 142.)

of pumping blood to the arteries, while if one sets off to explain doctors' use of the stethoscope in diagnostics, the contribution of the heart will be the noises it produces — something unlikely to be mentioned in an explanation of the circulatory system. This sort of perspectivalism is closer to the analytical account than it is to Craver's perspectivalism, falling thereby on the objectivist side of the debate about functions.

The point neo-mechanists want to make with this claim is akin to Cummins' point about different functional analyses falling from different analytical contexts. For the neo-mechanist, the same component can be part of several mechanisms at the same time, and may even have different systemic and/or teleological functions for each mechanism of which it is a part. A heart can be used as part of a device for tracking the passage of time, even while still pumping blood to the rest of the organism (by means of a sensor on the skin that counts heartbeats, for example). In this case, the heart has the function of pumping blood when the circulatory system of the organism is the mechanism of reference, while its function is that of beating regularly when the time-keeping device is the relevant mechanism. It does not follow that mechanisms or functions are observer-dependent. Both the circulatory system, and the time-keeping device may be objective features of the world — they simply share components. Perspectivalism about mechanistic explanations merely presses the point that depending on the mechanism of reference, the mechanistic and functional decomposition of a system may differ. As Piccinini (2015, p. 142) insists, this is an innocuous perspectivalism — one that does not threat objectivity. It signals a difference of explanatory focus, but does not threaten the objective nature of mechanisms, or functions.

In many cases involving biological systems or artefacts the different functions ascribed belong to two different types: one or a few teleological functions, many current or potential systemic functions. When the mechanism under consideration is the one that explains the mutual gravitational attraction between me and the Earth, my heart plays a role, has a function, insofar as it contributes to my mass, and therefore to my gravitational pull. However, such a mechanism is not a functional mechanism — it lacks any purpose, as much as the water-cycle, or the solar system. There is nothing that the mechanism behind my gravitational pull is supposed to do: it has no teleological function. The function of contributing to my mass that the heart has in such a mechanism is thus a systemic function, it is merely a causal contribution to a non-functional mechanism.

The case of the time-keeping device is more complicated. Let us suppose that the device has been designed by humans to work as a more or less precise clock[30]. In this case, the mechanism does seem teleofunctional: it has the function of keeping time to some degree of precision. Given the design of the device, it is arguable that the heart has the teleological function of beating sufficiently regularly. This is the contribution it makes to the teleological function of the overall mechanism. If the heart were to beat too irregularly, it would malfunction, it would hinder the time-keeping capacity of the

[30] I take this case to be akin to the case of diagnostic procedures with stethoscopes, *contra* Piccinini (2015, p. 103). However, it is arguable whether diagnostic procedures such as these qualify as mechanisms.

device. Contrast this with the previous example: there is no sense in which my heart would be malfunctioning if it contributed more or less mass in the mechanism explaining my gravitational pull. To push this point further, suppose that the time-keeping device was not designed, but evolved through natural selection — perhaps because keeping time to some precision proved to be advantageous to organisms in a particular habitat. It would seem odd to deny then that the heart does not also have the teleological function of beating regularly.

These examples illustrate that one can be a perspectivalist about mechanistic explanation, while being an objectivist about both teleological functions, and mechanisms. Perspectivalism about mechanistic explanation must be kept separate from perspectivalism about functions. While perspectivalism about functions entails perspectivalism about mechanisms, perspectivalism about mechanistic explanation entails neither perspectivalism about mechanisms, nor about functions. If the mechanism under investigation is the circulatory system, the heart has the function of pumping blood; if the mechanism under investigation is that of time-keeping, be it artificial or biological, the teleological function of the heart is beating regularly. Two different mechanisms, two different teleological functions contributed by the same component. No threat to the objectivity of either the mechanisms, or the functions follows.

As Piccinini (2015, p. 142) points out, objectivity would be in jeopardy only if, for the same target capacity, different and incompatible correct mechanistic explanations could be provided — if, for instance, there could be two correct and incompatible mechanistic explanations of the circulatory system. Therefore, the perspectivalism that is common in the neo-mechanist literature should not be confused with the considerably stronger, and non-innocuous type of perspectivalism put forward by Craver (2013). Given the close relationship between functional and mechanistic explanation, an anti-realist, perspectival view of functions leads to a non-innocuous perspectival view of mechanisms. Mechanisms, on this view, are ways of carving nature in more or less arbitrary ways according to our needs and interests. As Craver (2013, pp. 134-135) puts it,

> Mechanistic and functional descriptions ... presuppose a vantage point on the causal structure of the world, a stance taken by intentional creatures when they single out certain preferred behaviours as worthy of explanation... [functions] are imposed from without by creatures seeking to understand how a given phenomenon of interest is situated in the causal structure of the world.

This fundamental difference between the analytical account and strong perspectivalism nothwithstanding, when it comes to supplementing the mechanistic view of concrete computation with a notion of function, the latter fails for reasons not dissimilar to the ones that made the analytical account unsuitable for our purposes. Let us assess perspectivalism in light of the four tasks that the notion of function is supposed to fulfil in the mechanistic view of concrete computation.

The foregoing view of functions and mechanisms denies the objectivity of computa-

tion to which the mechanistic view of computation wants to cling, beside bringing back the spectre of pancomputationalism, which the computational mechanist is at pains to avoid. The traditional objection to Cummins' analytic view of functions presented in the last section also affects perspectivalism. By relying on current dispositions of components and systems for function ascription, it becomes mysterious how to make space for malfunction. One possible path to follow is to argue that the normativity of function is itself also imposed from without: when functional talk is appealed to in explanation, its normativity is, as it were, included in the package. The 'creatures seeking to understand' a certain phenomenon may already have criteria for what they consider as proper functioning as opposed to malfunctioning — they may already impose, that is to say, a view on what the system under investigation, as well as its components, are supposed to do. I will not pursue this line of reply, since it involves giving up objectivism about the normativity of functions, a consequence that is at odds with the aims of the mechanistic view of concrete computation.

Finally, as for the analytical account, perspectivalism does not square well with our pretheoretical intuitions about what kinds of systems in the world count as having teleological functions. Non-designed and non-biological systems, such as planetary systems or the water-cycle, can be analysed by means of the analytical strategy, yielding functional ascriptions to their components. This flouts our intuitions that functions apply exclusively to artefacts and biological systems.

In sum, perspectivalism is clearly not suitable for fulfilling the needs of the mechanistic view of computation: it fails to satisfy all four *desiderata* set out above.

Propensity theories

The analytical account and perspectivalism are not the only available declensions of dispositional theories of function. Propensity theories also have recourse to dispositions of components and systems in accounting for function. However, instead of embracing liberality of function, or making function hinge on explanatory interests and other pragmatic factors, propensity theories individuate one or more special capacities, or overall dispositions, that ground the possession of functions.

I will focus here on the version of the theory developed in Bigelow & Pargetter (1987), targeted specifically at biological functions. According to their proposal, functions are properties of organisms and their components that lead to a disposition that enhances their capacity to survive. Survival is the privileged capacity that bestows functions on biological systems and their components — only those dispositions that contribute, or would contribute, to survival count as functions. Taking survival as the relevant capacity for functional ascription is not supposed to depend on our explanatory interests, thereby avoiding the observer-relativity that marks the perspectivalist position. There are no competing choices of capacity that ground organisms and their parts possessing functions.

Dispositions to contribute to survival are relative to particular habitats, as Bigelow & Pargetter (1987) are quick to point out. Dispositions that enhance survival in one

environment may hinder it in others. How to determine the habitat against which to measure the survival value of a disposition? Bigelow & Pargetter (1987) admit that this is not straightforward, and help themselves to a vague notion of 'natural' or 'usual' habitat. It is against the backdrop of the natural habitat of a type of organism, or of internal components of organisms, that biological functions can be determined. Hence, the components of an organism that has always lived in a hostile, 'non-natural' environment, still have functions — for they would contribute to survival were the organism to live in its natural habitat.

Let us apply this theory to our running example, the heart. The heart has the function of pumping blood because in the natural habitat of organisms with hearts it is the disposition of hearts to pump blood that contributes to the survival of the organism. On the other hand, the disposition to produce throbbing noises does not enhance survival in the natural habitats of such organisms. Against this latter point, it may be argued that, in a habitat in which diagnostic procedures for cardiac health involving stethoscopes are widespread, the noises generated by the heart do contribute to survival. By allowing heart conditions to be discovered readily and treated, the disposition to produce throbbing noises contributes to the survival of the organism. It is at least arguable that there is a sense of 'natural habitat' that would include such diagnostic procedures. A case can be made that in contemporary societies those diagnostic procedures, given how widespread and ingrained in our cultures they are, make part of the natural habitat of humans as well as domesticated animals.

Much thereby hinges on how to define the notion of natural habitat. Too much vagueness risks opening the doors to an overly liberal notion of function (*e.g.* if the whole universe, or possible worlds could count as natural habitats). Nonetheless, even a generous notion of natural habitat, one that for instance makes most contemporary human societies count as such, can lead to a sufficiently delimited account of biological functions. It does not seem absurd to claim that in contemporary societies the throbbing noise of hearts does have the function of communicating the state of the heart to medical practitioners.

Bigelow & Pargetter (1987, p. 195) argue that the noises produced by hearts do not count as teleological functions. For, they argue, heart-noises would be present even if they made no contribution to survival. They are parasitic on the function that actually explains the presence of hearts in organisms, namely their capacity to pump blood — the noises being merely a by-product. I think, however, that Bigelow and Pargetter cannot appeal to this kind of considerations, on pain of abandoning the dispositional approach. Dispositional theories of functions are forward-looking — they are concerned with the 'how does it work' question, and not the 'why is it there' one. Heart-noises do contribute to survival in most contemporary human societies, insofar as they contribute to widespread diagnostic procedures. Whether or not this is why hearts are present in humans in most contemporary societies is a different question, one that has little to do with dispositions to contribute to survival. Much hinges on how to define 'natural habitats'. At any rate, it seems that, barring implausibly liberal

construals of the notion, the foregoing account still leads to an objective and non-trivial notion of biological function[31].

I will not purport to assess the general merits of propensity theories as theories of function[32]. My aim is much more limited, *i.e.* evaluating whether such theories can provide a notion of function that the mechanistic view of concrete computation can use. Let us then focus on the four requirements laid out above.

First, Bigelow and Pargetter's propensity theory seems to provide an objective notion of function, insofar as no recourse is made to observer-dependent properties. Though the notion of 'natural habitat' is vague, the theory still delivers, barring exceptionally generous construals of that notion, non-trivial functional ascriptions. Not all systems or their components would have the function of computing. Consequently, pan-computationalism is avoided: since computational systems are functional mechanisms with the teleological function of computing, and not all physical systems are functional mechanisms with such a function, not all physical systems are computational.

Lacking a general theory that also subsumes functions of artefacts, it is impossible to assess whether the theory accords with our pretheoretical intuitions. The scope of Bigelow and Pargetter's theory is limited to biological systems, one of the domains to which we tend to ascribe teleological functions. Whether it can be satisfactorily expanded to the other domain, designed systems, is unclear. This introduces a note of caution for the computational mechanist lured by the foregoing propensity theory.

Importantly, the theory fails the normativity requirement, for it has trouble with the notion of malfunction in a way similar to perspectivalism[33]. Bigelow and Pargetter's propensity theory entails that systems and components that we pretheoretically regard as dysfunctional do not have functions at all. For these lack dispositions to contribute to survival — given how they are constituted, they lack the disposition to do what functioning ones do. A severely malformed or damaged heart does not have the disposition to contribute to the survival of the organism, quite on the contrary. It follows from the foregoing account that such a heart lacks any function — it could not pump blood adequately in any circumstances, and therefore cannot and could not contribute to the survival of the organism. This is an unfortunate result. We tend to regard such cases as still involving functions: the said heart has the function of pumping blood, but fails to perform it[34]. Bigelow and Pargetter's propensity theory fails adequately to capture the normativity of function, an important requirement for the mechanistic view of concrete computation.

[31] A complicated task that a full theory should address is that of justifying why survival, rather than something else, should be the capacity of relevance for determining biological functions. Another issue is whether this kind of theory can be expanded to include not only biological functions, but also functions of artefacts. Bigelow & Pargetter (1987, p. 194) briefly hint at such an expansion, suggesting a general notion of function grounded on propensities for selection, be them natural or artificial. It is worth pointing out how this general theory seems to be at odds with the theory for biological functions just presented, which has no role for a notion of selection.

[32] For criticism, see Godfrey-Smith (1994, pp. 352-55.)

[33] Millikan (1989*b*), Neander (1991), Piccinini (2015).

[34] This is so even in cases in which a heart never pumped blood. Think about infants born with severe organ malformation — we still ascribe a function to the malformed organ, even though it never performed it, and never could have.

The propensity theory just examined, in sum, meets the first two requirements set by the computational mechanist on a satisfying theory of function; but it fails the third, and is silent on the fourth. In consequence, it seems unfit to complement the mechanistic view of concrete computation.

Goal-based theories

The final type of dispositional theory of function that I will examine, goal-based theories, have been proposed by, among others, Boorse (1976) and Piccinini (2015). According to these theories, teleological functions hinge on the goals of organisms and artefacts, the latter perhaps being derivative on the former. More precisely, functions are held to be contributions to the goals of organisms. A natural question arises: how to determine the goals of organisms? In the case of conscious, deliberating agents such as human beings, some goals are consciously entertained, intentions are had, and plans are made. But in the great majority of function ascriptions we are not dealing with conscious deliberating agents, even when we are concerned with human beings. The goal or goals that hearts contribute to need not, and generally are not, consciously entertained by organisms. Moreover, in the great majority of situations, hearts continue to perform their functions regardless of conscious goals individuals might have.

Boorse (1976) offers a goal-based theory in which the goals of organisms that help determine functions vary from context to context. Functions, he claims, are relative to system, goal, and time[35]. However, he gives no criteria for determining what the goals of systems are at any point in time, opening the door to pernicious liberality of function ascription. Moreover, making functions relative to time introduces further liberality. As Boorse admits, at a certain moment a luckily placed book may contribute to the goal of surviving by stopping a bullet from damaging vital organs. In this context, the book will have had the function of stopping the bullet, thereby preserving vital organs from damage, and contributing to the goal of the organism to survive.

This liberal, context-dependent account of functions is too generous in its function ascriptions to be of any use to the mechanistic view of concrete computation. Almost any physical system can be used at a certain point in time as a special-purpose computer — a falling stone may be used to compute the height of a tower, or the value of gravitational acceleration of the planet it is in. This would lead to limited pancomputationalism, since almost any system can have the teleological function of computing in specific circumstances. Though it arguably maintains the objectivity of functions, since no observer-dependent properties are appealed to by the account, Boorse's view is unsuitable to the mechanistic view of computation due to its liberality. Moreover, it is liable to the same line of objection regarding normativity that plagues pespectivalism and propensity theories.

Other goal-based theories are considerably less liberal, and hold promise as providers of a robust notion of teleological function that helps make sense of objective, non-trivial, and normative computation in physical systems. An alternative goal-based theory in

[35] Boorse (1976, p. 80.)

which liberality is curbed has been proposed by Piccinini (2015). He focuses on what he dubs objective goals of organisms: goals that organisms are organised in such a way as to strive to attain. One of these goals is survival, understood as the maintenance of processes and states that protect the organism from perturbations, and allow its continuation. Another objective goal is what Piccinini calls inclusive fitness: organisms are organised in such a way as to allow the production and protection of new organisms similar to themselves, either directly or indirectly (*e.g.* by protecting the offspring of a similar organism)[36]. Two factors justify the selection of survival and inclusive fitness as the objective goals of organisms: first, if those goals were not fulfilled, organisms would cease to exist, for they would be incapable of preserving themselves, and reproducing; second, organisms expend energy in striving to fulfil such goals.

Objective goals form the basis of Piccinini's theory of teleological functions:

> A teleological function in an organism is a stable contribution by a trait (or component, activity, property) of organisms belonging to a biological population to an objective goal of those organisms. (Piccinini 2015, p. 108)

Something analogous can be said of artefacts: their teleological functions are their contributions to the objective goals of the organisms that created them[37]. This applies to all or most artefacts created by animals, and at least to some artefacts designed by humans — such as clothing, and spears. These contributions need not be direct, as in those examples. Computers, for instance, contribute to complex analyses and planning, which in their turn improve production of goods that contribute to the survival and inclusive fitness of human beings (at least in theory).

However, many human-created artefacts do not contribute to survival and inclusive fitness — think of lava lamps, selfie-sticks, or Tetris — and a surprisingly large number of them actually hinder fulfilment of those goals — think of addictive recreational drugs, cigarettes, contraceptives. Nevertheless, these artefacts are generally taken to have functions, though it is arguable whether they have teleological functions, rather than merely systemic functions. At any rate, such devices do seem to have purposes and ends, they are designed for displaying a certain capacity, and can be said to be malfunctioning when they fail to do so.

To accommodate these cases, though remaining neutral on whether they really involve teleological functions, Piccinini introduces the notion of subjective goal[38]. Subjective goals are goals that sentient, and rational organisms may have other than the objective goals of survival and inclusive fitness. They may relate to pleasure, interest, avoidance of pain and boredom. A general theory of teleological functions takes into consideration contributions to objective and subjective goals of organisms.

Let us see whether Piccinini's goal-based theory succeeds in furnishing a suitable notion of teleological function to the mechanistic view of concrete computation, as it was intended to do.

[36] Piccinini (2015, p. 105-6.)
[37] Piccinini (2015, p. 111.)
[38] Piccinini (2015, p. 116.)

First, as with Boorse's account, the objectivity of teleological functions is guaranteed. No recourse to explanatory interests, or other observer-dependent properties is made. Second, defining what goals are, and determining which kinds of goal are relevant to the fixation of functions helps avoid pancomputationalism. The contribution that most components and artefacts make to objective (and subjective) goals are not computational in nature. Moreover, such contributions must be stable, and cannot be accidental happenings as in the case of the bullet-stopping book. Even though complex enough physical systems can be used to compute in specific circumstances, and thus contribute to some subjective goal, their contribution is not stable or frequent: it only takes place in very specific circumstances, while most often those systems contribute a different capacity to a different objective or subjective goal, if any[39]. Third, the foregoing account encompasses biological systems and artefacts, ascribing functions exclusively to them, as our pretheoretical intuitions would want.

However, there is the risk that normativity will prove tricky to this theory as much as it did to the ones analysed above: deformed or damaged components that do not make stable contributions to objective (or subjective) goals would seem excluded from possession of functions. It would follow that they have no function, rather than having a function that they fail to perform. As for the propensity theory, malfunction appears to be impossible, putting the normativity of function in jeopardy.

Piccinini's theory manages to dodge this problem, for the proposed definition of function makes reference to biological populations — and more generally to types, thus also covering functions of artefacts. Hearts in humans contribute to survival by pumping blood. A token heart that does not is a malfunctioning heart, and not a heart deprived of function, insofar as hearts in the overall population have the function of pumping blood. Similarly, a computer that does not compute is malfunctioning inasmuch as it belongs to a type that has the function of computing. In this way, the notion of malfunction, and the normativity of function, are preserved.

There is one unfortunate consequence that makes the normativity that comes out from this account clash with our pretheoretical function ascriptions[40]. Suppose that a mutation causes most tokens of an organ to stop working properly, and thus cease to contribute to the objective goals of organisms. By the lights of the foregoing theory, that organ would cease to have a function, for most tokens of the type would not contribute to survival and inclusive fitness[41]. This is at odds with our pretheoretical intuitions: if all kidneys were to cease filtering blood due to a viral epidemic, we would say that all kidneys became dysfunctional, not that they suddenly ceased to have a teleological function. Piccinini (2015, p. 114) hints at a possible way out, *i.e.* counting as tokens not only present instances of a type, but also past instances (within individuals). So kidneys could still retain their function given the fact that they did in the past — before the epidemic. This move introduces a historical element in the account, but only to a

[39] See Piccinini (2015, p. 112.)

[40] This point also applies to dispositional theories that appeal to statistical considerations, as versions of Boorse's, and Bigelow and Pargetter's theories.

[41] Neander (1991, pp. 182-83).

very limited extent. Selectional history is not brought to bear, but only the capacities that the same token, now diseased, had before the disease. Thereby the account can still claim to be largely ahistorical.

Counterintuitive results would still follow. Suppose further that the virus that diseased existing kidneys also generated a genetic mutation that ensured that the offspring of diseased organisms would also feature dysfunctional kidneys. At some point in the future, most kidneys would never have had the capacity to filter blood, and would thus lack teleological functions. To avoid this problematic consequence, a stronger appeal to historical considerations would be required, *e.g.* to the (deep or recent) selectional history that led to the existence and persistence of kidneys in the phylogeny. This move would turn the foregoing view into something much more alike historical, or backward-looking, theories of function, which I will examine in the next section. The goal-based theorist is thereby presented with a choice between a counterintuitive consequence of the account, which is though considerably limited in scope, and may arguably prompt a revision of the pretheoretical use of the term; or giving up the ahistoricity of the theory. However goal-based theorists may go on this issue, a robust notion of normativity is available.

Piccinini's goal-based theory respects all four requirements on a notion of function that can suitably help compose the mechanistic view of concrete computation — perhaps unsurprisingly, since it was tailored to do so. We therefore have the existence proof that we needed to defend the cogency of the mechanistic view: at least one available, plausible theory answers to the job description set out by the computational mechanist. Though this is a sufficient result, I will go on and analyse the other broad family of theories of function — based on the notion of selected effects — to examine whether it can provide notions of function that respect the four *desiderata* above as well.

5.3.2 Selected-effects theories

The main alternative to dispositional theories appeals to history in bestowing teleological functions. Selected-effects theories have recourse to the selectional history of a component or system in determining its functions[42]. This sort of theory takes the relevant question that appeal to function answers to be: why is a certain component or system there? The answer is given in terms of the capacities that past instances of the component or system displayed that explains the existence and persistence of instances today. The capacity that explains the existence and persistence of a component or system today is the teleological function of that component or system. An early version of a theory of this kind was put forward by Wright (1973). His proposal is the following:

> The function of X is Z means
>
> (a) X is there because it does Z,
>
> (b) Z is a consequence (or result) of X's being there. (Wright 1973, p. 161)

[42]This terminological choice is not without problems, since there can be dispositional accounts based on dispositions to be selected. Bigelow & Pargetter (1987) hint at such a view. I rely on the past tense of 'selected' to underline the historical dimension of the family of theories here examined.

According to this view, what a component does, the effects it has, explains why it is there. Hearts are there because they pump blood, and pumping blood is an effect of hearts being there. Pumping blood is therefore the function of hearts. This formulation is very general, and does not directly refer to selection processes, or to historical considerations. The way the effects of a component explain its presence in a system is left open. This generality invites crippling objections.

Boorse (1976) considers the case of a leak in a hose that contains a poisonous gas. The leak causes the release of the gas into the room, making so that anyone who tries to fix it falls unconscious due to inhalation of the gas. According to Wright's analysis, the leak would have the function of releasing the poisonous gas, for the leak persists because it releases the gas (knocking unconscious anyone who tries to fix it), and the release of the gas is a consequence of the leak. This shows how, without further constraints, the foregoing view is overly liberal: it ascribes functions in cases in which function ascription is inappropriate.

The best way to improve over the pernicious liberality that marks Wright's view is to offer more precise constraints on the processes underlying (a). In the case of biological functions and non-human artefacts, natural selection is the main, but not the only, candidate; while for human artefacts, intentions, as well as rational selection between alternatives may play a role. What kinds of processes underlie (a), and whether there is one type of such processes or many, are the main factors that tell apart different theories belonging to the selected-effects family.

The basic structure offered by Wright has been further developed in different ways by a host of theorists, among which Millikan (1984, 1989*b*), Neander (1991), Godfrey-Smith (1993, 1994), Buller (1998), Garson (2011). Selected-effects theories tend to be less heterogeneous than dispositional theories. Many of the differences between distinct views stem from the more or less general notion of selection employed, as well as the temporal span of the selection history relevant for function fixation (*e.g.* deep *vs.* recent history). I will dedicate little space to the varieties of selected-effects theories, since in many cases their differences are subtle, and of little relevance to my present purposes.

The paradigmatic kind of selection process appealed to in these theories is natural selection. A trait or organ has a function if it contributed to the reproduction of organisms having it, causing the trait or organ to survive in organisms across generations. Such contribution is its function. The reason why the trait or organ is (or was) present in organisms is its having done something in past organisms that increased their reproductive success, and thereby caused the reinstatiation of the trait or organ in new organisms in the same phylogeny. In some views, natural selection is taken to be the only kind of selection that bestows functions on physical (biological) systems, *e.g.* Neander (1991), Godfrey-Smith (1994). Other views tend to be more liberal, and allow other kinds of selection processes to play a role in bestowing functions. Nevertheless, it is undeniable that natural selection — and to a lesser extent sexual selection — have pride of place in selected-effects accounts of function. Natural selection putatively provides the means to account for most biological functions, with other forms of selection covering relatively

smaller domains, and sometimes being parasitic on it.

The weight given to natural selection, even in accounts more liberal than Neander's (1991), risks neglecting the important role that other selection processes may play in bestowing functions. As Garson (2011) points out, there are several cases in which function ascriptions are made in biology that do not involve phenomena for which natural selection is a plausible selection processes. Biologists are happy to assign functions to traits and behaviours that develop ontogenetically, such as neural and antibody selection, as well as some forms of learning[43]. Natural selection, being effective only phylogenetically, cannot account for these cases. Moreover, such traits get selected by means other than differential reproduction, on which natural selection is based. Accounts that rely heavily on differential reproduction may be overly restrictive, failing to cover some cases of function ascription in biology.

These considerations lead to the 'generalised selected-effects theory' proposed by Garson (2011, p. 555):

> The function of a trait consists in that activity that historically contributed to its being differentially reproduced or differentially retained within a biological system.

Insofar as it encompasses most of the factors that play important roles in different accounts, I will take Garson's generalised theory as representative of selected-effects theories of function. Its broadness is welcome, as it provides an unified account supposed to cover all biological functions. But there is a further factor that distinguishes selected-effects theories from each other, as already hinted. It hinges on how to understand 'historically' in the definition above.

Some views understand the appeal to history as an appeal to deep history, that is, to the contributions to differential reproduction and retention that took place in the remote evolutionary past of the phylogeny. This approach is more congenial to views that adopt a narrow understanding of what selection processes are relevant, inasmuch as deep history is not directly relevant to learning, and other selection processes that happen in ontogeny, while it is relevant to natural and sexual selection. Other views take the appeal to history to involve only recent history[44]. This position is compatible both with narrow understandings of selection processes, such as Godfrey-Smith's (1994), and with more liberal views, such as Garson's. Whereas in the former case the appeal is to recent evolutionary history, in the latter it also includes ontogenetic and learning histories of individual organisms.

Though I believe that views that rely on different kinds of selection processes, such as Garson's, are much more promising, I need not take sides. More relevant to my aims is to assess whether the notion of function that comes out of selected-theories accounts satisfy the requirements on theories of function that the mechanistic view of concrete computation puts forward.

[43] Garson (2011).
[44] Godfrey-Smith (1994).

5.3.3 Assessing selected-effects theories of function

Objectivity

Selected-effects theories seem quite easily to respect the objectivity requirement set by the mechanistic account of concrete computation. The selection processes they have recourse to are independent of the cognitive states, and explanatory practices of humans. All the selection processes appealed to — natural selection, neural and antibody selection, as well as some forms of learning — are objective features of the world, taking place independently of the presence of observers. By grounding function on such processes, selected-effects theories yield objective, observer-independent functions. Biological traits already had functions long before humans, or even sentient animals, existed. Relatedly, selected-effects theories base functions on purely naturalistic grounds, as the selection processes at work are *bona fide* naturalistic processes. The scientific credentials of the notion of function in the biological sciences are preserved.

Nevertheless, there are reasons for dissatisfaction regarding the way objectivity is obtained in selected-effects accounts. Objectors have insisted that this family of views produces a notion of function that has little explanatory value inasmuch as it involves only reference to past events, and not to the current causal dispositions of physical systems. As Piccinini (2015, p. 102) puts it: "the history of an organism's ancestors cannot contribute to the causal powers of that organism's components or properties". History, the claim goes, is irrelevant to the current causal powers of systems.

Take two identical connectionist networks. One has been trained to distinguish male from female faces, training that led it to have the weighted connexions it has; while the other was generated by a random process. Both have the power to distinguish male from female faces, and can therefore be used in face-recognition software. But selected-effects theories would have it that while the former has a function, having undergone a process of selection, the latter does not, as it has no such history[45]. This strikes many as counterintuitive, as it fails to do justice to the present causal powers of systems.

For related reasons, selected-effects theories are accused of making functions epiphenomenal. Since the function-bestowing selection processes happen in the past, they are not directly connected to the causal powers of current tokens of the functional type[46]. What function a system has is largely independent of its actual causal powers, being fixed by the causal powers of past tokens of the type which led to its selection. While this helps the selected-effects account to give an answer to the malfunction problem that plagues dispositional accounts — as we will see in more detail below — it does so by disjoining functions from current causal powers, thus arguably making the notion of function epiphenomenal.

At the heart of this dispute lies the more fundamental debate regarding what question the notion of function is supposed to answer. Proponents of dispositional theories take the relevant question to be something akin to: what causal contribution does the system make? On the other hand, friends of selected-effects theories take the question

[45] This is a less fanciful version of the 'swamp' argument. See also Eliasmith (2000).
[46] See Artiga & Martínez (2016).

to be answered to be: why is the system or component there?

The latter are unlikely to be swayed by the causal explanatory value and epiphenomenalism objections. For these objections only grab a hold when it is the former question the one considered to be central. There is no problem in answering a why-question without having recourse to current causal powers of systems. On the contrary, it is to be expected that questions about the reason why something is present will rely on historical factors. Moreover, selected-effects theories do provide a causal explanation in answering such why-questions, for the selection processes appealed to are causal processes. From the fact that there is no appeal to current causal powers it does not follow that the explanation provided is not causal. The bite of this line of objection to selected-effects theories hinges largely on the more basic dispute about what role, or roles, the notion of function plays is in our scientific pursuits. Unless this fundamental issue is solved, the objections from epiphenomenalism and lack of causal explanatory value lose much of their weight[47].

In sum, selected-effects theories can provide an objective, observer-independent, notion of function. They do so by means of an appeal to selection processes that took place in the past, partially disconnecting the notion from the current causal powers of physical systems. Though this has led to charges of epiphenomenalism, and lack of causal explanatory value, these only grab a hold once a stand on deeper issues about the role of the notion of function in science is adopted — a stand to which selected-effects theorists are hostile. In the current state of play, those objections fail to motivate the rejection of selected-effects theories, and the objectivity of function that ensues from it.

Selected-effects theories hence respect the objectivity requirement placed by the mechanistic view of concrete computation.

Pancomputationalism

Selected-effects theories also help avoid pancomputationalism. Only systems that have undergone selection processes (of one or more kinds) are candidates for possessing functions. Given that few types of processes qualify as selection processes, and that most physical systems do not undergo those process, it follows that most physical systems are not candidates for possessing functions. Since for the mechanistic view of concrete computation only systems that possess functions can be candidates for being computational, and given that selected-effects theories rule out most physical systems from possessing functions, it follows that most physical systems are not computational. Pancomputationalism is avoided.

As for propensity theories, selected-effects theories are often limited to biological functions, so that artefacts, such as artificial computers, are outside their purview. In extending a selected-effects theory to encompass functions of artefacts, two complementary paths seem to be available: (a) take artefactual function to be derivative of biological function, artefacts being selected by processes such as natural selection due

[47] Pluralism about function, I submit, is the most promising approach, though I will not argue for it here. See Tinbergen (1963), Milkowski (2016).

to their increasing the fitness and survival chances of the organisms that used them; (b) see artefactual function as stemming from a different kind of selection process, in which rational choice in light of the objectives and intentions of individuals or collectives selects the most adequate artefact or artefact design.

The former path seems to be most suitable for artefacts built by non-human species, such as beaver dams, ant mounds, and the sticks chimpanzees use to catch insects. The second path fits best human-built artefacts, which often go through processes of individual or collective planning. During planning, design proposals that are judged to be inferior to alternatives in light of set individual or collective goals are rejected, and superior ones are selected. In either case, it is to be seen whether the resulting accounts succeed in avoiding being too liberal in bestowing functions on physical systems, which might open the doors for a limited pancomputationalism restricted only to non-biological systems. I will not assess this possibility, as it would involve a lengthy diversion. Though a non-liberal selected-effects theory of artefactual functions does not seem particularly difficult to come by, it is worth keeping in mind that, as things stand, the risk of a limited pancomputationalism remains present. That notwithstanding, there are good *prima facie* reasons to believe that selected-effects theories have the tools to help the mechanistic view of concrete computation avoid pancomputationalism, while at the same time encompassing both biological and designed computational systems.

Normativity

Perhaps the crucial advantage of selected-effects theories in relation to dispositional theories is how they handle the normativity of function. As we have seen in section 5.3.1, dispositional theories have considerable difficulties in providing an appropriate notion of malfunction. Selected-effects theories, in contrast, have a ready way to account for the normativity of function.

Recall that dispositional accounts have trouble because they cannot easily disjoin the function of a system from its current causal powers, making it so that a system cannot have a function if it lacks the appropriate causal powers. Selected-effects theories avoid this difficulty insofar as they do not appeal to current causal powers in grounding the notion of function. A system has a function insofar as it or, more typically, its ancestors were selected because of the effects they had in the past. A malfunctioning system lacks the causal powers required to bring about the effects that led to it or its ancestors being selected. But it has a function insofar as it has an appropriate selection history, regardless of whether it is currently capable of performing it.

The appeal to past history of selection that characterises selected-effects theories gives them the tools adequately to respect the normativity of function, and in a way compatible with scientific naturalism. Functions are bestowed by selection history, regardless of the actual causal powers of token systems. In this way, systems that for one reason or another do not have the appropriate causal powers still retain their functions. They are malfunctioning insofar as they are incapable of bringing about the effects they were selected to bring about. A heart has the function of pumping blood if

in the history of selection hearts were selected because they pumped blood, leading to the instantiation of token hearts. If due to some congenital disease a heart is unable to pump blood, it is a malfunctioning heart — if fails to perform its function insofar as it is incapable of pumping blood, the effect for which hearts were selected, selection that explains the existence of that particular heart.

In sum, by severing the close tie between current causal powers and functions put in place by dispositional theories, selected-effects theories are able to account for the normativity of function, and preserve an adequate notion of malfunction — as the mechanistic view of concrete computation requires.

Intuitive appeal

Selected-effects theories, as we have seen, are mostly thought to account for functions of biological systems, one of the two kinds of physical systems which we tend pretheoretically to regard as teleofunctional. But unless theories of this family are broadened to encompass artefacts, they clash with our intuitions, insofar as they deny that artefacts have functions in the same way that biological systems do. At any rate, there are ways to broaden the scope of selected-effects theories so as to make them account for artefactual functions as well. If such attempts should succeed, as seems plausible, selected-effects theories would cover the two domains in which the common-sense notion of function grabs a hold, namely biological systems, and artefacts.

Two outcomes of selected-effects theories arguably clash with pretheoretical intuitions.

First, selected-effects theories deny functions to systems that have the appropriate causal powers to generate an effect, but which lack a selection history. Two identical systems, with identical causal powers, may differ in the functions they possess depending on whether they are the products of selection processes, or merely of chance. This result supposedly clashes with general intuitions about when function ascription is appropriate, and explanatorily helpful[48].

Second, first instances of a certain trait or system do not have functions. As they do not have appropriate selection histories, they have no function by the lights of selected-effects theories, even though their producing the effects that they do leads to selection and persistence of themselves, and perhaps of future instances. This outcome arguably goes against our pretheoretical intuitions, as we would normally be happy to assign functions to first instances.

Even admitting that selected-effects theories produce these counterintuitive results, arguments based on pretheoretical intuitions have limited weight. Measured against the theoretical virtues, and the explanatory fruitfulness of a theory, the alleged counterintuitiveness of some of its tenets should not motivate its rejection, as the history of science illustrates profusely. More important for selected-effects theories is to make sure that artefacts are also contemplated, given how useful and productive the notion of artefactual function is. Proponents of selected-effects theories may either broaden

[48] See Piccinini (2015, pp. 102-3.)

the scope of their theories, as suggested above, or else embrace some form of pluralism about function — *e.g.* by letting selection processes determine biological functions, and contributions to objective and subjective goals determine artefactual functions. In the latter case, it is crucial for the purposes of the mechanistic view of concrete computation that all the theories of function employed respect the four requirements set above.

5.4 Concluding remarks

An objective notion of teleological function is crucial for the mechanistic view of concrete computation to succeed in its attempt to overcome the shortcomings of rival theories. The reliance on such a notion is controversial, and potentially problematic, for the notion of function itself is to some extent an obscure concept awaiting a fully satisfactory philosophical treatment. As I showed in this chapter, the prospects for the mechanistic view of concrete computation are positive. There are plausible theories of function available that can play the role required by the mechanistic view, thereby securing its superiority in relation to competing views of concrete computation. In particular, two families of views, goal-based theories and broadened selected-effects theories, meet the set requirements.

The aim of this chapter was not to defend or endorse a particular theory of function, but was rather to provide a proof of existence: theories of function that have the characteristics required by the mechanistic view are in hand[49]. As I have shown, there are theories that are up to the task.

With the mechanistic view of concrete computation now thoroughly defended, it is time to get back to the problems that exercised us in Part I. It is time, that is, to come back to the issue of representation in the cognitive sciences, an issue that becomes considerably more tractable, I believe, once we are armed, as we presently are, with a sufficiently robust notion of concrete computation.

[49] Partly for this reason, I have not attempted to provide an exhaustive survey of theories of function put forward in the literature. Promising views that I have not tackled include Krohs' (2009) design-based account, endorsed by Milkowski (2013), and Mossio *et al.*'s (2009) organisational account. These views might prove to be as satisfying as the two that I have favoured in this chapter.

Part III

Deflating Content

Chapter 6

Interpretational Semantics

6.1 Taking stock

It is time to take stock. In previous chapters, I have introduced and analysed foundational issues in the cognitive sciences, involving the notions of computation and representation. Representation plays a central explanatory role in most of the cognitive sciences, and is frequently appealed to in the scientific literature. It is an essential theoretical posit and, as such, it is often taken for granted in the empirical study of the mind.

Philosophers interested in the foundations of cognitive science have worried about the conceptual underpinnings of the notion, its naturalistic status, and its explanatory role. As we have seen in Part I, the explanatory work that the notion of representation is supposed to carry out in empirical theories requires that any philosophical account of representation meet some requirements: it must explain how representations come to be contentful; how they come to have the contents that they do; what makes them into representations (Ramsey's job description challenge); and it must make space for the possibility of misrepresentation. The philosophical quest to provide answers to these challenges features a marked naturalistic leaning: all these requirements, the mainstream lore would want, should be met by having recourse exclusively to naturalistically acceptable entities and relations, so as to give the notion of representation a respectable place in the scientific worldview[1].

Part I of this work was dedicated to understanding the scientific importance of the notion of representation, and has explored the most influential attempts at providing a theory of representational content, together with the challenges they must overcome. These challenges take different forms, though most of them are underlain by the requirement of yielding fairly determinate representational contents, as well as to make space for misrepresentation, in order to do justice to the explanatory role of representation.

I focused especially on a particularly promising, though ancient, theory of representation: structural representation (Swoyer 1991, Gallistel 1990, Cummins 1989, 1996, Ramsey 2007). As we have seen, at the basis of such a theory lies the idea that representations act as stand-ins for things in the world, thus enabling what Swoyer dubs

[1] How to understand what counts as naturalistically acceptable is of course a vexed question.

'surrogative reasoning'. Though this theory has much to recommend it, as it fares reasonably well when confronted with the requirements on a satisfying theory of representation, it has a crucial shortcoming: it leads to wild non-uniqueness of representational content, a form of indeterminacy of content. Proposed fixes to non-uniqueness, *e.g.* by Bartels (2006), Isaac (2012), Shea (2014), are promising, but not free of difficulties.

The idea that the mind is, or is to be explained as, a computing system was one of the founding pillars of the field of cognitive science in the 1950s, and is still today at the basis of many research endeavours. Seeing the mind as a computing system has considerably improved our understanding of the workings of cognitive systems. Explaining what computing systems are, and how the mind can be said to be a computing system, or to be usefully explained as one, is another foundational issue in the cognitive sciences. Part II was concerned with the notion of concrete computation, or computation in physical systems. By having a good understanding of what it is for a physical system, such as a computer or a cognitive system, to compute, the hypothesis that computation plays an important role in cognition can be clarified and substantiated.

The notion of representation has traditionally come hand-in-hand with that of computation. A very popular position is expressed by Fodor's (1981) famous slogan 'no computation without representation'. Computation, according to this view, relies on semantically individuated states — it consists of a special kind of representation manipulation. I examined this position, and argued that it is problematic. Views of concrete computation that do not rely on semantic properties are more promising and should be preferred. I presented and assessed alternative, non-semantic views of computation, such as Chalmers' (2011) sophisticated causal mapping view and, especially, the rising mechanistic view (Piccinini 2007*b*, Milkowski 2013, Fresco 2014). I argued that the mechanistic view, when properly amended, is particularly promising. It is able to preserve the objectivity of computation, as well as its valuable explanatory purchase — suitably carving up the domain of physical computational systems — while doing justice to the practices of computer and cognitive scientists.

My aim in Part III is to provide a theory of representation and representational content able to play the required explanatory role in the cognitive sciences, while steering clear from the metaphysical difficulties to which existing theories of representation may be liable. My approach will be deflationary. I will rely on the robust notion of concrete computation provided by the mechanistic view to individuate computational structure as one of the factors that carries the most load in explaining complex appropriate behaviour. Ascription of determinate representational content comes on top of that, and depends on the task at hand, and on the context the organism finds itself in.

In this, I will be accepting the invitation made by Piccinini (2004) to conjoin existing theories of content, which have traditionally relied on a semantic view of computation, with a non-semantic view of computation instead. I take that structural representation is a particularly promising candidate for such a treatment. It is a notion of representation which arguably answers Ramsey's 'job description challenge'[2], and which is often

[2]Though see Morgan (2014).

at work in empirical research. In this chapter, I present and examine a relatively underrated theory of representation: interpretational semantics. This analysis is the first stepping stone building up to the deflationary view of representation I intend to defend. Interpretational semantics is closely related to structural representation, as I show in the next chapter, though it is considerably friendlier to the idea of doing away with robust reductionism about representation. By bringing the robust notion of computation provided by the mechanistic view to the fore, I propose a middle way between robust reductionism and full-blown pragmatism — a view of representation that populates the nearly empty grey zones between the extremes, merging interpretational semantics and structural representation into a hopefully more promising account.

I will begin with an analysis of interpretational semantics in its two main guises, one put forward by Cummins (1989), the other by Ramsey (2007). As we will see throughout the chapter, though Cummins' and Ramsey's accounts have many similarities, they are also importantly different. After briefly presenting interpretational semantics in section §6.2 and section §6.3, I turn in the following sections to examining some crucial elements of the account. I analyse and criticise the notions of computation and interpretation at play in the theory, which lead, in a way analogous to what happens in structural representation, to non-uniqueness of content. I conclude that, as it stands, interpretational semantics is unsatisfactory, though it provides some elements for a more promising theory of representation in cognitive systems, which I will put forward and defend in the following chapters.

6.2 Interpretational Semantics: preliminary considerations

Interpretational semantics was defended by Robert Cummins in his book *Meaning and Mental Representation* (1989), and developed in a somewhat different direction by Ramsey (2007), in his book *Representation Reconsidered*. Interestingly, while other theories of representation put forward in the 1980s, especially informational semantics and teleosemantics, enjoyed numerous adherents, and gave rise to intense debating and a flurry of papers and books, interpretational semantics ended up being a rather unpopular sibling. It has been picked up by few, though important works in the field (*e.g.* Stich 1992, Ramsey 2007), and has even been abandoned by its own creator (Cummins 1996).

Part of the reason for this relatively limited interest in the theory may be due to the clear delimitation of its scope. As Cummins (1989) repeatedly emphasises, interpretational semantics is concerned exclusively with the notion of representation at play in the Computational Theory of Cognition, which, according to him, is mostly silent on issues concerning propositional attitudes and folk psychological states, such as beliefs and desires. As in the foregoing, the focus is on providing philosophical foundations for the cognitive sciences, rather than trying to vindicate common-sense psychology, and everyday content-ascriptions. Some authors have even placed interpretational semantics in the same group as putative pragmatist or instrumentalist theories of representation,

such as Dennett's (*e.g.* Piccinini 2004)[3].

The fact that interpretational semantics neglects propositional attitudes in favour of representational phenomena that have been more closely tackled by the cognitive sciences is of course no reason to reject it, or to underestimate it. We can distinguish two types of representation, for which two different accounts may need to be given: representations as they figure in scientific theories of cognition, which generally focus on subpersonal states; and representations as they figure in the propositional attitudes, such as beliefs and desires. As I pointed out in Part I, there is no convincing reason to believe that one and the same theory of representation is going to be able to account for these two very different sorts of cognitive phenomena — which I labelled, respectively, 'representational content', and 'intentional content' [4]. On the contrary, once their dissimilarities are appreciated, just the opposite should be expected. Therefore, rather than being a shortcoming, the insistence of interpretational semantics in restricting its scope is laudable — it refrains from conflating two distinct phenomena, and treat them as if they were one and the same. This is, alas, the exception, rather than the rule in the literature on representation.

Most other theories of content on offer fail to draw the distinction, and either target propositional attitudes (*e.g.* Dretske 1981, Stich 1983, Block 1986, Papineau 1987), presupposing that giving an account of those exhausts the problem of mental representation; or attempt to give a unified account supposed to account for both representational and intentional content (*e.g.* Millikan 1984, Dretske 1988). Often these theorists have developed theories of content that were clearly of the representational kind, rather than the intentional, and used them as stepping stones to providing an account of beliefs and desires. At any rate, I suppose that the closer relationship these theories have to the propositional attitudes, a traditional subject of philosophy of mind, is partly what has earned them so much interest, leaving interpretational semantics mostly in the shadow.

The project that motivates the foregoing work is clearly on Cummins' side. As I stated in Part I, I am here concerned with the philosophy of the cognitive sciences, rather than with the philosophy of mind, if this distinction has any meaning to it. As a consequence, the features that might have diverted the attention of philosophers toward other theories of representation should hold little or no sway over us presently. Our focus here is on representational content, not on the intentional content of propositional attitudes.

Computational cognitive science is a rich and advancing research field. Despite some opposition from radical embodied cognition researchers (*e.g.* Van Gelder 1995, Hutto & Myin 2013), most of cognitive science employs the computational framework with considerable success. Some of the objections against computationalism, such as Searle's (1980), target a stronger thesis, namely that the computational framework can exhaustively explain intentionality, propositional attitudes, and consciousness. Computational

[3]Though Dennett's early views were indeed instrumentalist, this cannot be said of his subsequent work.

[4]Cummins (1989).

cognitive science need not endorse any of those explanatory aims[5]. Interpretational semantics, or IS for short, is an interesting, though perhaps underdeveloped theory of representation that is meant to capture the explanatory role of representation in the cognitive sciences. Let us now turn to the view itself, as well as its merits and demerits.

6.3 Interpretational Semantics: the theory

As for structural representation, at the core of interpretational semantics lies the idea that representations stand in for what they represent. More precisely, representation is seen as falling out of what Cummins calls 'simulations'. Simulations take place when there is an interpretation of computational processes over computational objects that maps computational processes and objects into processes and entities in the world — that is, when there is a mapping of states and processes of the computational system onto entities and their patterns of transformation in a certain domain. In this way, those computational states and processes come to be representations of the entities and their transformations onto which they are mapped.

According to Cummins, the first step for a physical system to acquire representational properties is for it to compute a certain function *f*. In its turn, to compute a function f is to execute a programme that gives the appropriate outputs given inputs; that is to say, if $f(i) = o$, then the programme will produce output o on input i[6]. Programme execution is understood in terms of going through the steps of an algorithm, each step being the computation of a function, leading, step by step, from i to o. A physical system satisfies a function g when the causal goings-on in the system can be mapped onto the algorithmic steps in g[7].

The second step for a physical system to represent is that there be an interpretation mapping between the arguments and values of the function g that it satisfies, and objects and processes in a different domain. These processes and objects need not be cognitive in nature. By satisfying g, the physical system instantiates function f over those objects and processes to which the interpretation maps its arguments and values. In the best scenario, the mapping is an isomorphism — there is a perfect one-to-one mapping between g and f. The interpretation mapping, which for Cummins is purely mathematical in nature, bestows on the elements and processes of g a representational nature — they represent the elements and processes in f to which they are mapped.

To use Cummins' example, if g is a function isomorphic to the adding function $+$, then the arguments and values of g can be mapped onto the arguments and values of $+$. In this way, the physical system comes to be seen as an adding machine. As a consequence, its inputs, outputs, and internal operations take on a representational

[5] On the limits of the computational framework when applied to these sorts of cognitive phenomena, see Stich (1983), Fodor (2000).

[6] Cummins (1989, p. 91.)

[7] Systems may satisfy functions without computing them (*e.g.* Venus satisfies the Newtonian equations). This poses a problem for IS as presented by Cummins. He does not provide a criterion for deciding which systems compute and which do not, if not for his problematic appeal to programme execution. I will come back to this issue below, in section §6.4.

character. Inputs and outputs become representations of numbers, and the internal operations of the system become manipulations of numerical representations. As Cummins (1989, p. 93) puts it, "representation is just a name for the relation induced by the interpretation mapping between the elements of g and the elements of $+$". In other words, it is because the physical system mirrors or simulates, under a certain interpretation I, a function in another domain that its elements and processes acquire a representational coating. They become representations insofar as the system can be interpreted, given the isomorphism between its elements and relations, and the elements and relations of another domain, as simulating the latter. Importantly, that there be an interpretation mapping between the computational physical system and another domain is sufficient for bestowing representational status on the elements and processes of the former.

In the case of cognitive representations, the interpretation mapping assigns to elements and processes in cognitive systems the entities and relations that they represent in another domain. As such, the interpretation mapping reveals the computational processes going on in the cognitive system as realising a cognitive capacity; it provides a linkage between the instantiated function, or 'mere state crunching', and cognition[8].

The task or cognitive capacity is the *explanandum* of cognitive science, and as such it is the starting point for theorising — it is given. The central explanatory role of representation in cognitive science, according to IS, is that of explaining how come certain physical/computational goings-on are able to give rise to adequate behaviour, or to a specific cognitive capacity. By means of the interpretation mapping, those goings-on reveal themselves as simulating the target domain or, as Ramsey puts it, as models of the target domain. As parts of a model, those physical/computational elements and processes acquire representational status — they stand in for things and goings-on in the world.

This account has been dubbed by Cummins (1989), quite fittingly, 'the Tower Bridge picture'. The interpretation mapping I works as the towers in the famous bridge in tying together what goes on in the bottom span, the computational processes in the physical system, and the top span, consisting of the function or cognitive ability under analysis. Since by means of the isomorphism the elements and processes in the bottom span track elements and processes in the top span, the physical system, by satisfying a function g over its states, can be interpreted as simulating a function f over the tracked target domain. In interpretational semantics, simulation under I has primary importance, with representation coming as an instantaneous consequence. The primary role of simulation in the resulting account led Cummins to dub this kind of representation 'S-representation', where the 's' stands for 'simulation'[9].

Interpretational semantics and its Tower Bridge picture apply not only to cognitive representations and cognitive abilities. It also works for representations in non-cognitive systems, such as designed calculators, as well as for scientific representation. In the case of the latter, we have objects and processes in one domain, for instance geometry, being mapped onto, and thus representing, objects and processes in other domains, for in-

[8]Cummins (1989, pp. 110-1.) See also Egan (2010).
[9]Cummins (1989, p. 97.)

stance physical processes and quantities, such as velocity and distance[10]. Consequently, the notion of representation that falls out from IS is deflationist *sensu* Burge (2010): it fails to be a notion of exclusive interest to cognitive science, for there is nothing intrinsically *mental* about representation — the notion is also appropriate to domains that do not concern cognitive states[11].

In sum, interpretational semantics is a theory of representation designed to fit the explanatory practices and needs of computational cognitive science, though it also illuminates the notion of representation in non-cognitive domains. It is based on the idea that representation takes place when there is an interpretation which maps elements and processes of a computational system onto elements and processes of another domain. By means of this mapping, the computational system simulates the goings-on in the other domain. Simulation is at the core of the account of representation offered. It is thanks to simulation, enabled by the interpretation mapping, that elements and processes in a physical system are seen as stand-ins, as representations, of elements and processes of a different domain. By behaving in the physical system in a way functionally analogous to the way things in a different domain behave, elements of the computational system acquire representational status.

As an aside, a *caveat*: Ramsey (2007, pp. 102-4) convincingly argues that Cummins fails to distinguish two different kinds of representation in computational cognitive science. The Tower Bridge picture, according to Ramsey, captures not S-representation, but rather a notion that he dubs 'IO-representation' (from input-output representation). The latter is only concerned with the inputs and outputs of the subsystems that a functional decomposition of the overall system reveals. In order for these subsystems to be seen as contributing to the capacity featured by the whole system, it is required that their inputs and outputs be interpreted in light of the overall capacity. For instance, a system capable of multiplication may have as one of its components a subsystem that instantiates a certain function which, in order to explain its contribution to the capacity to multiply featured by the overall system, needs to be interpreted as taking numbers as input and delivering (repeatedly added) numbers as output. The subsystem's inputs and outputs are thus interpreted as representations of numbers, for this is explanatorily useful in accounting for the system's overall capacity to multiply. Arguably, no simulation need be involved in this kind of representation, though Ramsey admits that it is likely that most cases of IO-representation (especially those involving mathematical functions) are also cases of S-representation. In sum, while IO-representations represent by virtue of the role they play in explaining the capacity of the system to which they contribute, S-representation represents by simulating, or modelling, some target domain.

Regardless of whether the distinction turns out to be actually significant, I have tried to introduce the notion of S-representation in interpretational semantics in such a way as to keep it distinct from the IO notion, by stressing the importance that there be an isomorphism between the function satisfied by the physical system, and a function

[10] See Cummins (1989, pp. 94-5.)

[11] For more on this sense of 'deflationism', see chapter 8 below.

over the target domain. In my treatment of IS I will be mostly concerned with this 'corrected' understanding of S-representation.

Ramsey (2007) has endorsed most of the tenets of interpretational semantics in his analysis of current projects in cognitive science, and of whether they retain and make use of a robust notion of representation. His version of the theory features some differences to Cummins' formulation, which are worth spelling out in some detail. I will take a closer look at some of the crucial notions at work in IS, and show how Cummins and Ramsey understand them, where they agree, and where they differ. This perusal will allow a better assessment of the advantages and shortcomings of interpretational semantics, in its two guises.

6.4 Computation

The notions of computation, representation, and interpretation are enmeshed in an intricate tangle in interpretational semantics. In this section, I will briefly examine the notion of concrete computation that underlies Cummins' and Ramsey's view, and in the following sections I will turn to the other two notions. I will start with Ramsey (2007).

Due to the nature of Ramsey's (2007) project, his remarks on concrete computation are quite scant. His main concern is not to endorse one particular theory of cognition (*e.g.*, classical computationalism *vs.* connectionism), or of representation over another. Rather, his interests lie in enquiring on which extant theories of cognition actually posit a substantial notion of representation. His conclusion is that only an amended classical computationalist approach, and the interpretational semantics that falls from it really posit *bona fide* representations. Ramsey (2007) is mostly non-judgemental on whether we should keep to the notion of representation, and thus to the classical computationalist approach, or adopt eliminativism — though he sees the cognitive sciences moving toward the latter option.

Ramsey endorses the view of concrete computation that has pride of place in classical computationalism in cognitive science — the semantic view. As he recognises, most philosophers in this tradition have passed over going into much detail about what it is to implement a computation. They have generally endorsed the semantic view as the most natural one, as it seems straightforwardly to allow the conceptual transition from taking the mind to be a computational system to taking it to be a representational one[12] — concrete computation involving, according to the semantic view, the manipulation of contentful symbols.

Cummins has written more extensively on computation, but the picture he offers, though non-semantic, is congenial to the classical computational one (Cummins 1983, 1989, Cummins & Schwarz 1991). As we have seen in 6.3, Cummins takes computational processes largely to involve physical processes to which a systematic interpretation in terms of semantic properties can be provided[13]. However, according to Cummins

[12] Ramsey (2007, pp. 38ff.)
[13] Cummins (1983, p. 34.)

computation *per se* need not involve representational interpretation. Though computational processes in computational systems are normally carried out over objects with semantic properties — symbols — having semantic properties is not a necessary condition for computation. The view of computation that Cummins defends is thereby non-semantic, though he is quite willing to come to terms with proponents of the semantic view. Cummins is ready to accept that in the most interesting cases of computation — those involving designed computers and cognitive systems — computational processes typically involve representations[14].

At the core of Cummins' view of computation lies the notion of programme execution: to compute is to execute a programme (Cummins 1983, 1989, Cummins & Schwarz 1991). Computation is programme execution, and programme execution, in its turn, is understood as ordered step satisfaction: the programme is analysed into a sequence or network of component functions which are carried out in the appropriate order[15]. A physical system implements a programme when there is a mapping between the arguments and values of the functions called for by the specification of the programme, and the causal structure of the system. For Cummins, "functions satisfied by [a physical device] d specify causal connections between events in d", and "a system executes the program if that causal network [specified by the programme] gives the (or a) causal structure of the system"[16]. We have thereby a version of the causal mapping view of concrete computation, analysed in its Chalmersian version in section 3.3.1.

Let us get back to the example of the calculator computing the addition function $+$. According to interpretational semantics, $+$ cannot be directly computed by a physical system, insofar as it involves abstract entities, such as numbers. The only way for a physical system to compute $+$ is to satisfy a function g from physical inputs to physical outputs that can be mapped onto $+$. An adding machine computes addition when its causal structure, its satisfying g, can be interpreted as carrying out the steps involved in the addition function — whereby the elements and processes of the machine become representations of numbers and of numerical operations[17].

There are many problems with this view, for reasons familiar to us from Part II.

First, equating computation with implementing a programme either excludes from having computational status systems, such as finite state automata or connectionist networks, that are not normally seen as executing programmes; or blurs important distinctions in computer science by making programme execution ubiquitous[18]. Cummins' view of programme execution as step satisfaction leads to the latter horn of the dilemma. Finite state automata, connectionist networks, and other computational systems execute programmes on this generous understanding of programme execution — they do satisfy functions that count as steps in algorithms. Therefore, they count as computational systems; a welcome result. This liberal notion of programme execution

[14] See Cummins & Schwarz (1991, p. 62.)
[15] Cummins (1989, pp. 93-4.)
[16] Cummins (1989, p. 92.)
[17] Cummins (1983, pp. 36, 42.)
[18] Piccinini (2015, pp. 14-15.)

allows Cummins' view to capture the domain of computational systems[19].

However, as Piccinini (2015) points out, such a liberal notion does not fit well with distinctions that are made in computer science between different kinds of computational systems. Computers that have and execute stored programmes (in the traditional, non-liberal sense) are seen as more flexible than computational systems, such as finite state automata and connectionist networks, that lack stored programmes.

In contrast to Piccinini, I do not think that there is much weight to this objection from descriptive adequacy. What seems to be doing the job in distinguishing more flexible from less flexible computational systems is not programme execution *per se*, but rather the presence or absence of *stored* programmes. Finite state automata and connectionist networks are less flexible not because they do not execute programmes, but rather because they can only execute one programme (unless they are reprogrammed). They cannot switch to other programmes stored in memory, as can stored-programme computers. What is crucial for the flexibility of *stored*-programme computers is not programme execution, but rather their capacity to store several programmes in memory, and employ them as need and user dictate. Cummins' view does make space for the relevant distinction, *contra* Piccinini.

A second objection from descriptive adequacy moved by Piccinini is, I believe, more successful, though far from decisive. Piccinini (2015, pp. 157-8) notes that Cummins' account does not consider primitive computational devices, such as AND-gates, to be computational, for their operations cannot be divided into further function-computing steps. Since primitive computational devices do not execute programmes — insofar as their workings cannot be explained in terms of step satisfaction — they do not count as performing computations by Cummins' lights. This flies in the face of the practices of computer science, in which primitive computing devices such as logic gates are routinely claimed to compute (logical) functions.

It is debatable, however, whether the mechanistic view that Piccinini advocates is superior to Cummins' on this regard. For Piccinini, logic gates, when not part of an enclosing computational system, compute only trivially[20]. They cannot be explained computationally insofar as their behaviour cannot be broken down into more primitive computations. Explanation is here purely mechanistic, not computational — *i.e.* involving physical properties, such as voltage levels, rather than digits and operations on digits[21]. But when part of an enclosing computational system, Piccinini argues, primitive computing devices become computational insofar as they acquire the function of computing a certain logical function in the context of the whole system[22]. The mechanistic view has it that logic gates, when components of a computational mechanism, compute. Thus at least to some extent the mechanistic view is more descriptively adequate than Cummins'.

Descriptive adequacy considerations are at any rate of secondary importance, and

[19] Though it fails to capture exclusively the domain of computational systems, as I will show presently.
[20] Piccinini (2015, p. 156.)
[21] Piccinini (2015, p. 155.)
[22] Piccinini (2015, p. 156.)

may be compensated by theoretical virtues elsewhere. The crucial argument against Cummins' view of concrete computation derives not from his claim that computation is programme execution, but rather from his claim that step satisfaction is a matter of causal mapping between physical system and abstract computational description. This claim makes Cummins' view fall prey to trivialisation arguments as much as Chalmers' causal mapping account.

I briefly rehearse here some of the arguments already moved against Chalmers (1995, 2011) in section 3.3.1.

First, no criterion for how to group physical states, and their causal relations is offered by Cummins. This opens the doors to charges of triviality, since it is always in principle possible to group the physical states, and the causal relations of a complex enough physical system in such a way as to make them be mappable into any sequence of algorithmic steps. It follows that the causal goings-on in any complex enough physical system can be mapped onto potentially any programme, making performing concrete computations a trivial matter.

Second, as for Chalmers, limited pancomputationalism ensues. If having a certain causal structure is all that is needed to execute a certain programme, as Cummins (1989), Cummins & Schwarz (1991) claim, then any physical system with a causal structure executes a specific programme. Moreover, given the many different levels of causal description to which any physical system is subject, the same physical system executes different programmes at the same time. Even if a criterion for grouping physical states and processes should be provided, thus avoiding the unrestricted pancomputationalism that undermined the simple mapping view of computation, Cummins' account still entails limited pancomputationalism, making concrete computation into something nearly ubiquitous. This is of course a problematic outcome for a theory of concrete computation, insofar as we fail to drive a useful wedge between physical systems that are computational, and those that are not.

Cummins (1989) and Cummins & Schwarz (1991) are well aware of this problem. Simple causal mapping, they readily admit, leads to triviality[23]. They try to steer clear from that outcome, but their solutions are far from satisfactory.

Cummins & Schwarz (1991) claim that the causal structure of a physical system is distinct from its computational structure, if it has any. As they admit, "there are lots of causal processes, and only some of them are instances of function computation"[24]. They hold that algorithm (or programme) execution provides the means to distinguish computational from causal structure — only some causal processes count as computing a function, namely those that follow an algorithm. However, since executing an algorithm is, by their lights, a matter of step satisfaction, which in its turn is a matter of causal mapping, they fail to offer a principled way to distinguish mere causal processes from processes that are, in addition, computational. Every sequence of causal processes can be interpreted as carrying out the steps in an algorithm. Limited pancomputationalism still ensues.

[23]See for instance Cummins (1989, pp. 102ff.), and Cummins & Schwarz (1991, pp. 63-64.)
[24]Cummins & Schwarz (1991, p. 63.)

A further attempt made by Cummins (1989) and Cummins & Schwarz (1991) is to appeal to semantic interpretability. Only those causal structures that can be interpreted as representations count as computational. This move draws the view considerably closer to semantic accounts of concrete computation[25]. A semantic constraint is added to the simple causal mapping view: causal mappings that reveal computations are those that allow representational interpretations of the goings-on in the computational system.

There is a clear tension here. Insistence on the non-semantic nature of concrete computation is often followed by reliance on semantic interpretation in order to individuate computations. For instance, Cummins & Schwarz (1991, p. 62) claim that "the objects of computation needn't be representations of any sort"; just to go on and say on the following page that a computational explanation of a device's capacity to multiply must start by identifying representational states internal to the device, and the algorithms defined over them[26]. Though the matter is far from clear[27], it would seem that a semantic constraint comes in only given an explanatory target, which leads to semantic interpretation of the objects of computation in terms of the target explanatory domain, but is not necessary for computation *per se*.

Even if we take the foregoing view to involve a semantic constraint, the appeal to interpretation is nonetheless problematic. Without a properly constrained notion of semantic interpretation, states and processes of a physical system can always be interpreted semantically, and in consequence, transitions between those processes can be seen as function computations. Lacking a robust notion of semantic interpretation, the processes internal to any physical system are compatible with an indeterminate number of different interpretations, and thus compute — simulate — many different functions. It follows that implementing a programme is trivial, and so by Cummins' lights, concrete computation is trivial[28]. This is so, at least, if no robust account of interpretation is given, which would allow one to distinguish acceptable, computation-bestowing interpretations from those that are not. In order to make sense of the notion of concrete computation at play in interpretational semantics we need closely to investigate the notion of interpretation. This will come as no surprise, I am sure. We will turn to that in the next section.

For now, note that the need for a robust notion of interpretation falls from the weak view of computation that is offered. It is because the causal mapping view is remarkably liberal that interpretation must be invoked in order to avoid triviality charges. The notion of concrete computation with which Cummins is working in his treatment of interpretational semantics leaves gaps that further factors must fill. In the next section, I will show that interpretation, as Cummins (1989) understands it, is not an adequate fix. Ramsey's (2007) version is more promising, though not free from problems.

In the coming chapter, I will argue that if interpretational semantics is to aspire

[25] See section 3.3.2.

[26] Perhaps they have in mind a hybrid view, such as Rescorla's (2013), though this is not made explicit.

[27] Indeed Milkowski (2013) considers the view put forward by Cummins & Schwarz (1991) to be a version of the semantic account of concrete computation.

[28] Milkowski (2013, n. 4, p. 204) presses a similar point.

to become a good theory of representation and representational content, an alternative notion of computation must ground it. I will propose that the mechanistic view of computation is the one most suited to the task. It allows IS to embrace its deflationary tendencies on content in a more satisfying way, giving rise to a theory more robust than extant versions. For, I argue, the explanatory role of representation and representational content, as well as the objective nature of their substrates are preserved. But so much for anticipation. Let us tackle without further delay the notion of interpretation in interpretational semantics.

6.5 Interpretation

I have argued that Cummins' appeal to interpretation is partly due to the insufficiently robust notion of computation which underlies his account. His invocation of the notion of interpretation is motivated, I take it, by the worry that only interpretation mappings, with their accompanying ascription of representational content, can make determinate the computations performed by physical systems. The hidden assumption is that computational structure cannot be pinned down if not by assigning representational status to the elements and processes of physical systems. Even though Cummins' view is non-semantic, being a kind of causal mapping account, it still requires that states be semantically individuated when applying the computational framework to the mind (as well as to other domains)[29].

In Cummins (1989), the notion of interpretation at work is merely mathematical — it consists of a mapping of the causal goings-on in the physical system into the arguments, values and steps of a function *f*, and the processes that allow its computation. Even in the simple, non-cognitive case of instantiation of mathematical functions, representational contents, *i.e.* numbers, are ascribed to the elements of the physical system. In this basic case, the elements of the physical device are ascribed mathematical content[30].

As Cummins readily admits, further constraints are needed. A simple-mapping view of interpretation leads to function instantiation being trivial. As previous chapters have shown, the easy availability of one-to-one mappings makes concrete computation into a trivial matter, and representational content into a wildly non-unique affair. Given that there is a mapping between addition and multiplication, any adding machine would be interpretable as a multiplication machine as well. There is a mapping between the elements and processes in the physical system and the addition function $+$, which would make the physical system into an adding machine. But there is also a simple mapping between $+$ and the multiplication function $*$, and thus, given the transitivity of one-to-one mappings, the physical system also instantiates $*$. If interpretation is a matter of one-to-one mappings, then adding machines are also multiplication machines (and much else, since other mappings can be found)[31]. Taking interpretation to be purely

[29] See Cummins & Schwarz (1991).
[30] For a similar view, see Egan (2010).
[31] See Cummins (1989, pp. 102-3.)

a matter of simple mapping is a non-starter: any physical system would instantiate all sorts of different functions, making function instantiation, and concrete computation, trivial.

Relation-preserving isomorphism is a step toward less liberality, as interpretations must additionally preserve the relationships between the elements and processes in the physical system in mapping them into the arguments and values of *f*. An instantiation of + would not be interpretable as an instantiation of $*$, given that addition and multiplication are not isomorphic. Though this move helps, it is not enough, according to Cummins. There are still an infinity of functions that an adder could be interpreted as instantiating. The adding function + is isomorphic to functions such as $2(y + z)$. Given the transitivity of isomorphism, the elements and processes in the physical system are also isomorphic to $2(y + z)$, whereby the system can be interpreted as computing that function as well (and all the other isomorphic functions)[32]. Even when taken to involve relation-preserving mappings, the notion of interpretation is still much too liberal.

Cummins recognises that these notions of interpretation will not do. He also recognises that he has no solution to the problem. He hopes that a principled account of what he dubs 'direct interpretation' will be forthcoming. Direct interpretation would include principled ways to avoid trivialisation, with requirements on the lines of: not having the instantiated function be already included in the interpretation, in which case all the work would be done by the interpretation rather than by the structure of the physical system; and that the causal goings-on in the physical system not be ignored by the interpretation mapping[33]. Admittedly, without a satisfactory notion of direct interpretation, Cummins' version of interpretational semantics is in deep trouble. If interpretations are (almost) trivial, any system simulates an indefinite number of domains, causing IS to become an useless theory of representation for the cognitive sciences. If simulations are trivial, explanations of behaviour in terms of them are always possible for any minimally complex system, from pebbles to walls, computers, brains, and galaxies.

Moreover, lacking a robust notion of direct interpretation, the view of concrete computation offered by Cummins falls victim to limited pancomputationalism, as we have seen above. As the discussion in Part II suggests, it is extremely difficult to fend off triviality objections by relying exclusively on some form of mapping. Even Chalmers' (2011) demanding requirement that the abstract computation be mapped onto the causal topology of the system leads to limited pancomputationalism. If we should want to adopt that demanding requirement, the problem of triviality of IS would not go away. If pebbles and walls are to be seen as implementing computations, albeit uninteresting ones, that will suffice to allow interpretation, and thus ascription of representational properties by interpretational semantics' lights (barring a robust and principled notion of direct interpretation). Moreover, if a pebble or a wall compute some function, they also compute all functions isomorphic to it, and, as uninteresting as they might be,

[32]Cummins (1989, p. 103.)

[33]Cummins (1989, p. 104.) Thus formulated, these constraints are admittedly not satisfying, whence the 'on the lines of'. As Cummins points out, there are ways of getting around them (*e.g.*, by making the interpretation consist of a look-up table).

there will be multiple possible interpretations leading to multiple instantiated functions. Even pebbles would be victims of the liberality of Cummins' version of interpretational semantics.

The core notion that Cummins' version of IS requires to get off the ground, *i.e.* direct interpretation, is a promissory note, and one unlikely ever to be paid. But what about Ramsey's story? It is time to see whether the notion of interpretation in Ramsey's IS, which is considerably richer than Cummins', is able to bear the required theoretical load.

In Ramsey's picture a physical system implements a simulation, and its elements and processes hence acquire representational status, only under the interpretation(s) which are explanatorily apt, given the cognitive capacity under investigation. The *explanandum* determines which interpretation is the relevant one: the one that maps states and processes of the cognitive system onto the phenomenon that scientific research has determined as the one in need of explanation.

Put this way, it may seem that Ramsey is proposing a pragmatist view of representation. After all, he is appealing to the explanatory purposes of scientific investigation in order to fix the content-determining interpretation. This may raise the suspicion that the content-determining interpretation is chosen by us, on grounds of explanatory usefulness. Representation and content would depend on other intentional states, the goals and aims of cognitive scientists.

This is not, however, what Ramsey has in mind. On his account, the explanatorily relevant interpretation is determined by how the cognitive system uses the (computational) structure. The embeddedness of the cognitive system in an environment, its organismic needs and physical capacities, come to inform the cognitive task of relevance — the one we are interested in explaining. In the context of a specific task, the cognitive system makes use of the structure as a model, as a simulation, of the target domain with which it is presently at grips. Even though that same structure can be interpreted as simulating all sorts of different target domains, only one of those interpretations is relevant for the organism's interaction with its environment in a given context. Only this interpretation, grounded in the use of the structure by the cognitive system, endows its elements and processes with representational content. All other possible interpretations, which map the structure at hand into other target domains, are irrelevant, and play no role in content-ascription.

In other words, the use to which the structure is put by the cognitive system fixes its target, *i.e.* what it is simulating. And, as the story goes in interpretational semantics, it is the fact that a structure is simulating a target domain that makes its elements and processes acquire representational content. It is in virtue of being part of a structure that simulates, or models, something else, that the elements of the structure are mapped onto entities and relations in the target domain, thereby becoming representations of those entities and relations. In brief: representational content is fixed by an exploited second-order resemblance between representational vehicle and target domain[34].

[34]For a similar picture, see Shea (2014). Ramsey (2007, p. 95, and passim), as is usual in the literature, talks of isomorphism. But, as already noted previously, this is a somewhat imprecise use of

The explanatory burden shifts thereby to the use a structure is put to by the cognitive system. Use is constrained by the embeddedness of the organism in an environment — its needs, goals and causal interactions with its surroundings. Embeddedness, thus understood, reveals the cognitive task the organism is performing: *e.g.*, navigating a terrain in order to pick food at a certain location and bringing it back to the hive; or escaping from potential predators to a safe place. The task determines the target of the simulation the cognitive system carries out in order to generate appropriate behaviour. The structure comes by this means to be endowed with representational properties. In Ramsey's words: "the content of S-representation can be fixed by the target of the model, and the target of the model is fixed by the cognitive activity we want explained ... [which] is typically dependent upon the way the system is currently and causally engaged in the world"[35]. A satisfying account of the use of a structure is thus called for. I will postpone discussion of this issue to section 6.6.2.

6.6 Non-uniqueness of content

6.6.1 Cummins on non-uniqueness of content

Recall that according to liberal versions of structural representation, defended for instance by Cummins (1996), representations represent everything that they structurally resemble. Given the liberality of structural resemblance, representational vehicles structurally resemble many different entities in the world. It follows that representations have wildly non-unique contents, jeopardising the explanatory value of appeal to representation, as well as the possibility of a useful notion of misrepresentation.

A similar dynamics is at work in interpretational semantics. By basing representation on second-order resemblance — on which the notion of simulation is grounded — IS leads, in a way analogous to structural representation, to wild non-uniqueness of content. Furthermore, the appeal to interpretation mappings composes the problem. For not only collections of vehicles represent anything they structurally resemble, but interpretations may map elements and processes of the physical system into functions over a target domain for which there is no perfect resemblance[36]. Unless, at any rate, principled constraints on which interpretations are acceptable are provided.

The problems of IS with non-uniqueness of content seem as serious as those that structural representation has to face. This is not surprising, as both views rely on second-order resemblance — non-uniqueness of content being a consequence of the reliance on such a liberal notion. The version of interpretational semantics championed by Cummins (1989) is at odds with the one defended by Ramsey (2007) on this regard. While Cummins embraces the view that S-representations have non-unique contents, Ramsey tries to provide the tools to make their contents unique.

---

the term, since it is generally meant to cover less strict forms of second-order resemblance as well.

[35] Ramsey (2007, p. 96.)

[36] This is also analogous to a problem with structural representation. Since perfect structural mappings are too much to ask of representations, on pain of making representational correctness extremely rare, structural representationalists must appeal to some form of approximate, less-than-perfect structural mapping. See Shea (2014).

For Cummins, S-representation must always be relativised to a target by means of an interpretation mapping[37], but the availability of different interpretations makes representational content cheap[38]. Cummins holds that interpretational semantics leads to wild non-uniqueness of content because a structure S "s-represents the arguments and values of any function it simulates"[39]. In other words, S represents everything it can be interpreted as representing, it simulates every target domain to which the function g that it satisfies is (to some degree) isomorphic. Such an approach to S-representation leads to wild non-uniqueness of representational content, given the liberality of interpretation mappings — especially in the absence of a robust notion of direct interpretation.

Any function g satisfied by a physical system can be given different appropriate interpretations. That is to say, any function g can be mapped, by a suitable interpretation I, to all sorts of different functions over different target domains. In a scenario that should be familiar from Part I, the multiplicity of available interpretations leads to non-uniqueness of representational content. Under different interpretations, the physical system simulates different target domains, and thus its elements take on different representational contents. Since, by Cummins' lights, representational content is determined by every interpretation, we end up with wild non-uniqueness of content.

Cummins (1989) does not take the liberality that follows from his account to be pernicious to representational explanation in the cognitive sciences. He claims that, in the end, all that matters for cognitive science is that there be a notion of correct or incorrect representation given an interpretation I. The strategy is to try and save the explanatory adequacy of the theory by appealing to the ways the notions of simulation and representation are used in the sciences. The main role of representation is to provide a conceptual connexion between what goes on in the physical system, and a target capacity, be it purely computational (*e.g.*, computing addition), or cognitive (*e.g.* generating behaviour appropriate to a certain situation). The fact that explanation is carried out with a target *explanandum* in sight — in the typical case, a cognitive capacity, or successful complex behaviour — narrows down which interpretation mappings I are explanatorily relevant.

It does not matter that the physical system under scrutiny satisfies a function g which can be mapped by means of some I_{n} to other target domains. All that the cognitive scientist needs care for is that, under a fixed interpretation I_a, the physical system instantiates a function that simulates the capacity of interest (*e.g.* navigating Paris). All the other interpretation mappings are explanatorily irrelevant. The fact that that same function g could be mapped by a suitable interpretation to a simulation of navigating Tianducheng — a Chinese city designed to copy the spatial layout of Paris[40] — is silent on how the system was able to successfully navigate Paris. Tianducheng would come into the picture only in case the explanatory purpose were that of explaining the system's capacity to navigate Tianducheng, or perhaps its capacity to navigate both

[37] Cummins (1989, pp. 92-3, 101.)
[38] Cummins (1989, p. 106, and passim.)
[39] Cummins (1989, p. 136.)
[40] I am assuming, unrealistically, that the spatial layout of Tianducheng is identical, or at least very similar, to that of Paris.

Paris and Tianducheng.

In this way, misrepresentation is arguably accommodated by interpretational semantics. Representation and misrepresentation are always relative to an explanatory target (Cummins 1989, p. 101). It is only with respect to the function that I picks out as the one simulated that it is possible to assess representational correctness. Thus my map of Berlin, when interpreted as a map of Paris, is a case of misrepresentation. Moreover, though the choice of explanatory targets is partly dependent on the domain of investigation of the particular scientific subfield, it is not unconstrained. As Ramsey (2007, p. 95) underlines, in the case of organisms capable of cognition, their embeddedness in an environment, their causal exchanges with their surroundings, their needs and goals, help delimit the nature of the cognitive capacities that call for explanation[41].

Once we have a reasonably good idea of what the capacity to be explained is, cognitive science looks for its (possible) physical/computational realisers, and ties back the relevant entities and processes to the target capacity by means of a suitable interpretation mapping. If this endeavour is successful, and an appropriate mapping is found, an explanation of how the organism behaved appropriately in a given situation is made available: internal states and processes simulate, or model, the target domain — S-representation is at work.

The hypothesised physical/computational entities and processes may fail to sustain the proposed mapping — if there is no (good enough) mapping between the functions to be connected by the interpretation, the hypothesis that one system simulates the other will have to be reconsidered. The proposed entities and processes will fail to represent the target domain, even though, under another interpretation, they may successfully represent other target domains. What matters is that, under an interpretation, the simulation is not accurate, making the elements and processes of the simulation misrepresentations.

Cummins' account seems *prima facie* to have the tools to keep the explanatory value of representation and misrepresentation in explaining cognition unscathed by the liberality of simulation, and of the consequent non-uniqueness of content. The guiding idea is to embrace non-uniqueness of content, but focus on specific interpretations when assessing representational correctness and error. However, there are problems with this move.

It is unclear how to home in on the relevant interpretations — the ones under which to assess representational correctness and error, and thereby adequacy and inadequacy of behaviour. A possible strategy is to rely on the aims and interests of cognitive scientists, who individuate the relevant interpretations in light of their explanatory purposes, and perhaps other pragmatic considerations. This line partially abandons the naturalistic project in explaining representation and content, and will be further explored in section §8.1.

The difficulty of picking out the relevant interpretation for each case is rendered more acute by the failure, admitted by Cummins, of providing a satisfactory, naturalistically

[41] Note though that on Ramsey's own view these epistemological considerations play a secondary role. See next subsection.

acceptable account of direct interpretation, as we have seen in section §6.5. Without such a notion, systems instantiate functions trivially, thus composing the problem of non-uniqueness of content — not only any structure can be interpreted as representing many entities, but any structure can be interpreted as representing potentially any entity or set of entities, given that there are no criteria on what counts as an acceptable interpretation. It is doubtful that an account of direct interpretation can be provided without having recourse to factors that would endanger the naturalistic status of the theory.

Furthermore, Cummins' version of interpretational semantics *de facto* eliminates the role played by the interpretation mapping in the account. To claim that S-representations have as their contents the appropriate entities in every possible target domain simulated by the system is to claim that what matters for content-fixation is merely second-order resemblance. Structures would simulate everything to which they are resemblant, making the appeal to the notion of interpretation mapping largely idle. Or else, and this is the path I believe Cummins takes, the interpretation mapping is being taken as the isomorphism itself. The appeal to interpretation is an attempt to determine the functional structure of the physical system that is relevant for computational ascription. Without the needed notion of direct interpretation, however, this attempt does not succeed. We are therefore left with a liberal appeal to isomorphism, analogous to the liberal appeal to structural resemblance typical of structural representation.

In sum, Cummins' approach to interpretational semantics is extremely liberal about representational status and representational content. But interpretational semantics need not follow Cummins' liberal view. Content can be fixed by one or few interpretations, provided that a naturalistic means of selecting them is provided. This is the strategy preferred by Ramsey (2007), to which we now turn.

6.6.2 Ramsey on non-uniqueness of content

For Ramsey (2007), content is not non-unique. Representations arise only thanks to a particular interpretation mapping, understood not as merely mathematical, but rather in terms of the use to which a representation is put. Even though there generally are different possible interpretations, or uses, that map a computational structure to different target domains, content-ascription is relative to each interpretation. Under an interpretation, representational contents are fairly determinate. Given the constraints on the interpretation of interest, motivated by the organism's embeddedness in the environment and, derivatively, the purposes of the explanatory endeavour, interpretation mappings tend to be fairly unique — only one interpretation is going to be relevant for content-fixation.

Interpretational semantics, under this reading, can avoid wild non-uniqueness of content because second-order resemblance is not doing all the content-fixing work — as happens, in contrast, in Cummins' version of the account, as well as on pure structural representation. There is, so to speak, a crucial difference in timing between Cummins' and Ramsey's versions of the theory. The explanatory interpretation mapping in Cum-

mins' IS comes, as it were, after, or on top of, content-fixation — representations are already endowed with (wildly non-unique) contents by then. In the foregoing approach to IS, on the other hand, it is one particular interpretation mapping that helps give representational status to the elements and processes of the cognitive system, and thence endow them with content — the one dictated by the use the organism makes of the cognitive structure in dealing with its current situation[42].

Since each interpretation normally maps elements and processes of the system onto unique entities and processes in the world, representational content is not non-unique. All possible interpretations taken together, content in IS would be as non-unique as in structural representation — the same structure can be used, with reasonable degrees of success, in many different situations. However, there is no reason, in IS, to consider all interpretations together (*contra* Cummins 1989) — typically only one (or a few) are relevant, and thereby effective in determining representational content. In consequence, non-uniqueness of content would be avoided.

As representational content gets fixed only under one interpretation at a time, and as Ramsey (2007) urges, only one interpretation (or a few) is generally relevant to explaining a cognitive capacity, the fact that there are other possible interpretations available — which ascribe different contents to the same vehicle — should not impress us. However, this comes at a price: the notion of interpretation at work is much richer than the purely mathematical one (morphism) used by Cummins. And while a purely mathematical notion of interpretation can safely be part of a naturalistic account of representation, a more loaded version may put at risk the naturalistic standing of the theory.

For Ramsey, interpretation is based on the use to which a certain computational structure is put by the cognitive system. That use helps, together with second-order resemblance, determine representational content. Out of the many mapping relations that cognitive structures sustain, only one is relevant for the use the organism makes of it. This is the one that determines representational content. Therefore, representational content is typically unique, and misrepresentation can be accounted for. Cases in which representations are used in a way that is inadequate to the task at hand count as instances of misrepresentation.

Use selects the relevant resemblance relation, thus fixing what the computational structure is a simulation of, and consequently what its elements and processes are mapped onto — what their representational content is. By having recourse to cognitive use, Ramsey keeps at bay the worry that his view might be merely instrumentalist, or that representation is an observer-dependent affair, or an artefact of our explanatory interests. His constant appeal to explanatory considerations is of an epistemological sort, rather than metaphysical — it concerns the problem of how theorists can get a grasp on what the cognitive task at hand is, and thus what the target of the simulation is supposed to be. Once we know (or have a good guess of) what the simulation is targeted at, we know (or have a good guess of) what the cognitive system is representing[43].

[42] Ramsey (2007, pp. 95-6.)

[43] This constant appeal to explanatory considerations has led Morgan (2014, p. 224) to hold that in

Ramsey's strategy is, I believe, promising, and it keeps interpretational semantics firmly wedded to the naturalistic project, eschewing any role for pragmatic factors in content-fixation. However, the notion of use to which he appeals in cashing out a richer notion of interpretation risks reintroducing traditional worries about naturalistic theories of content, which we have examined in section §1.5.

First, in order for the notion of cognitive use to help pin down the content-fixing interpretation mapping, we need a principled, naturalistic way of individuating cognitive tasks, and what counts as success (or failure). Teleosemantics may be a candidate for providing these conditions[44]. As we have seen, there are potential problems with such a move. Functional indeterminacy may creep in, making unambiguous task individuation, as well as success conditions for its performance, difficult to get in a principled way. Second, even if a naturalistic way of uniquely individuating cognitive tasks were at hand, Burge's (2010) objections to teleosemantics may then come into force. Given that behavioural success is arguably independent from representational correctness, appropriate behavioural use seems to be an inadequate criterion for helping fix representational content. A representation may be put to successful behavioural use by the cognitive system, even though it is inaccurate.

In spite of these challenges, I think that versions of interpretational semantics vaguely in the spirit of Ramsey's hold promise.

In a more satisfactory version of the theory, I take, teleology should not be directly connected to content-fixation as it is in teleosemantics. Using teleological functions to help determine what the tasks facing organisms are, and what would count as success, need not involve acceptance of teleosemantics. We need not, that is, endorse the view that teleological functions are the main factor in content-fixation. Rather, indeterminacy can be resolved once we admit to the set of content-determining factors a host of other considerations, some of which of a non-teleological nature. We should moreover accept that, in some cases, if not quite often, our incapacity to take into account that host of factors may make content ascription epistemically undecidable. These epistemic limitations would not impact negatively on our cognitive sciences, I take. These considerations will be further explored soon, as they lie at the basis of the mild reductionist view of representational content that I will put forward in section §8.2.

Another way of individuating cognitive tasks and their success conditions involves pragmatic factors. As scientists, or in our everyday dealings with the world, we are interested in explaining some pieces of behaviour as instances of performance of some previously identified cognitive task; where the identification depends on our explanatory interests and aims, or on our pretheoretical intuitions. We can then assess representational use, and thus correctness, in light of its enabling successful (or else) accomplishment of the pragmatically-individuated task. This abandons the naturalistic project,

Ramsey's view representational content ends up being 'radically observer-dependent'. I disagree with this interpretation of Ramsey, as I believe that the cognitive use of a structure has pride of place in his picture of content-fixation, something that Morgan (2014) seems partially to recognise.

[44] Indeed, Millikan occasionally sees teleosemantics as simply providing the normative dimension to a mapping- or resemblance-based theory of content (*e.g.* Millikan 2004). See Shea (2013*a*) for critical discussion.

yielding a version of interpretational semantics that makes use of a notion of interpretation cashed out in pragmatic terms. This view is very close in spirit to the pragmatic deflationary path I will examine in section §8.1.

In sum, Ramsey's version of interpretational semantics calls for further elaboration, and points to some interesting directions in which theories of representation and content can be developed. One may, as Ramsey himself, try to cling to the naturalistic project, providing perhaps a theory in which teleology plays a role, though not the one teleosemantics would want. Alternatively, one can partially reject the naturalistic project, and embrace content pragmatism about cognitive representations. These considerations will lead me to explore, in chapter 8, two flavours of deflationism about representation and content.

6.7 Concluding remarks

Interpretational semantics is a promising proposal for understanding the explanatory role played by representation and representational content in the cognitive sciences. Despite the critical attitude I took toward the view in this chapter, I am, as will become clearer in following chapters, largely sympathetic to its approach to representation and content. I think, however, that IS mistakenly tries to play a game to which it is not suited, namely that of naturalising content in robust reductionist terms. Interpretational semantics is better employed in the deflationary project that I will defend in the next chapters. In order for it to shed its robustly reductionist metaphysical chains, and at the same time keep its explanatory power, I argue that IS must embrace a stronger notion of concrete computation. I believe indeed that many of the issues with interpretational semantics stem from the weak and unsatisfactory notion of concrete computation that it employs. In the next chapter, I will conjoin IS with the mechanistic view of concrete computation. This move brings IS and structural representation close together. Most importantly, such an amendment makes space for a notion of representation that is deflationary in its metaphysical commitments, while nonetheless keeping its explanatory power in the cognitive sciences.

Chapter 7

Mechanising Interpretational Semantics

Interpretational semantics is a promising framework for understanding representational content in the cognitive sciences (short of beliefs and desires). However, extant versions of the theory feature shortcomings that cannot be ignored. The notion of interpretation is ill-defined, and at any rate insufficient to avoid triviality of computation; and non-uniqueness of content looms, as well as the risk of falling prey to some of the traditional arguments against theories of content. I believe that the problems with IS have an identifiable origin: the employment of an unsatisfactory notion of concrete computation. It is partly due to the weak notion of computation at play that appeal to interpretation and representation becomes necessary to individuate computational structure. Endorsing a stronger notion of concrete computation, such as the one provided by the mechanistic view explored in Part II, makes it possible to explore an alternative explanatory route: using concrete computation, mechanistically understood, to make sense of representation.

Metatheoretical considerations recommend such a path. A satisfying theory of representation has eluded philosophers despite the decades (or centuries[1]) of focused efforts. It looks like the notion of representation is a bad candidate for being part of the explanatory basis of phenomena such as concrete computation. Hence, if we are to take seriously the proposal that the mind is a computing system, it might be worthwhile to investigate alternative theoretical paths — paths that do not have as their starting point the notion of representation, but that may perhaps lead to it. In particular, I will enquire into whether the mechanistic view of concrete computation can help provide a satisfying theory of representation.

The bases for my proposal have been laid out in previous chapters. It is though worthwhile to go into more detail on the workings of the picture I am putting forward, so as to see how the many topics we have touched upon so far mesh together to provide what I take to be a satisfying theory of representation for the cognitive sciences. This will involve some review and repetition, which I hope will nonetheless be useful to the

[1] Or millennia.

reader.

7.1 Starting with mechanism

Let me briefly recapitulate some of the features of the mechanistic view of concrete computation that will be relevant for the coming discussion. For a more detailed treatment, I refer the reader to Part II.

First, recall that the states and processes of computational mechanisms are individuated non-semantically[2]. They are not individuated by their representational contents, but instead by their functions, *i.e.* the roles they play in enabling the mechanism of which they are part to have computational capacities — capacities that in their turn are one of the teleological functions of the system. A (digital) computational capacity, on its turn, is not a matter of manipulating semantically interpretable symbols, but rather of processing sequences of input, output, and intermediate strings of digits according to a set of rules sensitive only to some degrees of freedom of the vehicles.

Second, the mechanistic view does not rely on interpretation (neither mathematical nor cognitive) in order to characterise a mechanism as computational. If the system is a mechanism that has the teleological function of sporting computational capacities, understood as the manipulation of strings of digits according to rules, then the system is computational. The problems that surfaced above with the notions of interpretation and computation in interpretational semantics do not apply here. In opposition to Cummins (1989), direct interpretation plays no role, let alone an 'absolutely central' one[3], in concrete computation.

Some degree of something similar to interpretation is required. Insofar as a physical system normally features many different capacities, mechanistic explanations can be provided for each of them. For instance, laptop computers have capacities such as to heat up, to provide light, to make noises, etc. Mechanistic explanations for these capacities carve the components and processes of the mechanism in different ways. For instance, the electrical goings-on in the processor unit play the role of dissipating energy in the form of heat. This is an innocuous perspectivalism, as we have seen in chapter 4. The fact that a physical system may have many different capacities is unsurprising, as is the fact that those capacities require different explanations. What matters is that once the computational capacities of the system are the ones in focus, only one mechanistic explanation of those capacities is correct.

Embracing the view that computational systems are teleofunctional mechanisms deflects the risk of triviality, or excessive liberality of concrete computation. Since computational mechanisms have as their teleological function that of computing, their computing capacity is the relevant one. The other capacities the system can be said to have are parasitic, or secondary — provided, of course, that they are not additional teleological functions of the mechanism. Those capacities are not what the mechanism is for and, as such, may depend on interpretation to come to surface. The teleological

[2] Piccinini (2008*b*).
[3] Cummins (1989, p. 105.)

function of the mechanism, on the other hand, is independent of interpretation, as it is fixed by the objective processes that determine bestowal of teleological functions to physical systems according to the suitable theories of function that we have examined in section §5.3. Nothing detracts from the objectivity of computational mechanisms and of computations.

Third, the mechanistic view succeeds in avoiding limited, and *a fortiori*, unlimited pancomputationalism. It drives an appropriate wedge between physical computational systems, and physical systems such as walls, pebbles, and planetary systems, which, though computationally describable, do not qualify as computational. The latter systems are not bounded mechanisms that have computational capacities as we have here defined them: their components and processes do not involve digits, and there is no ordered manipulation of digits according to rules. Crucially, they lack the appropriate teleological function that marks out computational physical systems. Computational descriptions of such systems are possible, but that does not make them into computational mechanisms. The triviality objections against the notion of computation, which feed on the threat of pancomputationalism, do not take a hold[4].

These three features of the mechanistic view of concrete computation give us the initial tools for informing a theory of representation. Concrete computation, being non-semantically individuated, does not involve representation. It is thus in a good position to ground an account of representation[5]. The fact that concrete computation is objective makes the computational structure of cognitive states something objective, out there in the world rather than in the eye of the beholder. Finally, the mechanistic view is not vulnerable to the Putnam-Searle triviality objections. Concrete computation is not only objective, but it is also non-trivial. Most physical systems are not computational systems, and computational systems do not compute any possible function. It follows from the mechanistic view that a computational system typically has uniquely definable computational structures — the structures responsible for enabling the overall computational capacities of the system[6].

In brief, the mechanistic view of concrete computation provides a robust notion of computational structure. Computational structures are non-semantically individuated, objective (observer-independent), and determinate. With this notion of computational structure in our hands, it is time to reassess structural representation, as well as its relationships with interpretational semantics.

7.2 Mechanistic computation and Structural Representation

The robust notion of computation and thereby of computational structure that the mechanistic view of concrete computation provides plays a crucial role in the account

[4] See Piccinini (2007*b*).

[5] Piccinini (2004, 2008*b*).

[6] At least insofar as the argument against multiplicity of computations presented in section §4.4 succeeds.

I want to put forward. The picture involves complementing structural representation with the notion of computational structure mechanistically individuated — a move that brings structural representation and interpretational semantics close together.

My proposal is that we go back to the core insight of interpretational semantics, and especially of structural representation, that systems of representational vehicles enable successful complex behaviour by instantiating the relational structure of a target domain. I propose to complement that insight in two ways that are at odds with extant versions of IS and structural representation: first, a) I claim that the structures that play the explanatory representational role are computational structures mechanistically individuated; second, b) I argue that representation is less central for a theory of cognition than generally thought — computational structure carries most of the explanatory burden, with the notion of content having its metaphysical importance downsized.

The first claim distances the foregoing picture from interpretational semantics, and brings it closer to structural representation. The second claim, on the other hand, does the opposite. For a) posits interpretation-independent structures and processes that play the explanatory role of representations, giving us, in contrast to Cummins' IS, a principled and objective way of carving up the cognitive system, as well as well-defined computational and representational vehicles. Meanwhile, b) rejects the robust reductionism about representational content that underlies structural representation. In my view, the internal states leading to successful complex behaviour are not primarily individuated by their representational contents, but rather by their computational structure. I will come back to b) in the next chapter.

Structural representation is based on the idea that representations represent by virtue of instantiating the same relational structure, *i.e.* by being structurally resemblant, to what they represent. I propose that the proper way of cashing out the relevant relational structure of systems of representational vehicles is in terms of their computational structure. Collections of vehicles represent the entities in the world that share their computationally-individuated structure. Importantly, this introduces strong constraints on what internal states are candidates for representational status, and it also provides a definite way of individuating the relational structure that helps establish the content-fixing second-order resemblance relation.

First, only states that have a computational structure mechanistically individuated may attain representational status. This is a demanding requirement. The structure, to count as computational, must be such that it plays a role in the overall computational capacities of the cognitive system, such as, for example, computing the route back to the hive. This rules out structures that do not contribute in the appropriate way to the computational capacities of the whole system. For instance, while a battery in a laptop plays a role in enabling the system to compute insofar as it provides the system with energy, the role it plays is not itself computational. It does not carry out manipulations of strings of digits that are part of the overall computational capacity. Rather, it contributes by providing energy to those components of the laptop that do carry out computations (as well as to other noncomputational components, such as the

fan and the screen). In the cognitive case, structures formed by blood vessels, glia cells, neuronal organelles — and *ad hoc* hodgepodge structures — are excluded. Though these entities play vital roles in the cognitive system, they do not play a computational role (to the best of my knowledge[7]).

Viewing the cognitive system as a computational mechanism, as mainstream cognitive science does, allows the non-semantic individuation of its relevant functional structure. The elements and processes of the system are carved according to their computational roles[8]. It follows that the cognitive system has objective computational structures, *i.e.* relational structures over those of its elements and processes that play a computational role, and stand in computational relations to each other. Describing this structure is an empirical matter: it is an aim for the empirical branches of cognitive science.

Neurons, assemblies of neurons, and their activities — the realisers of computational and representational vehicles in the brain — stand in many sorts of physical relations to other parts of the cognitive system, many (or most) of which have no bearing on the cognitive capacities of organisms. Appeal to computations mechanistically individuated helps constrain the physical properties and relations that are relevant for cognitive abilities. Without such constraint, we lose sight of which subset of the complex physical goings-on in the system are cognitively relevant.

It is unlikely that looking for a full description of the computational structures of the whole cognitive system is going to be a feasible scientific enterprise, at least in the near future. More promising in the short-term is an approach that focuses on specific capacities of organisms, and tries to individuate the computational structures that enable it. The cognitive system most likely involves nested mechanisms — computational structures in the cognitive system have components that are themselves computational mechanisms, and whose computational structures may, on their turn, have components that are also computational mechanisms, and so on. This will bottom-down in primitive computing components[9], such as logic gates are for computers. In the case of the brain, it is less clear what the primitive computing mechanisms are. They might be realised by the activity of single neurons, or even single dendritic trees[10].

Philosophy cannot decide these empirical matters (though it can likely help). At any rate, the cognitive sciences are still far from having the tools to make hypotheses regarding the computational mechanisms, the computational structures, and the computational primitives of the cognitive system that are not highly speculative[11]. What

[7] Glia cells are perhaps the best candidates for having a yet-to-be-discovered computational role.

[8] There may be cognitive abilities that exploit in part non-computational processes, such as random processes. The foregoing account does not rule this out, but rather focuses on those abilities, arguably the majority, that either in part or fully employ computational processes. Note moreover that the definition of generic computation offered in section §4.2 leaves open the possibility that some cognitive processes compute analogically.

[9] See Piccinini (2007*b*, p. 510.)

[10] See for instance Ryder (2004), who develops a theory of representation based on dendritic computation, and second-order resemblance. It is a good illustration of one of the shapes the philosophical picture I am here advocating may come to take if future empirical work vindicates it. I do not buy Ryder's theory of content though, as will become clear.

[11] But some headway has been made. See, for a review, Carandini & Heeger (2012), and section §8.4

philosophy can do, at this point at least, is to enquire on whether the foregoing framework has good prospects of solving issues in the philosophical foundations of the cognitive sciences. The mechanistic view of computation is a big step ahead in the foundational project, insofar as it provides an objective, robust account of concrete computation. I believe that, in its turn, this robust notion of concrete computation helps with one of the other foundational, and deeply problematic, notions in the cognitive sciences — representation.

The mechanistic view of concrete computation provides a definite, objective notion of concrete computation, and of computational structure. By itself, this does not tell us anything about representation, since computation and computational structure are non-semantically individuated, and need not be interpreted as playing a part in any kind of model or simulation. However, this view of computational structure can be integrated in a theory of representation for cognitive science.

In the theory I am putting forward here, computational structure comes to take the place of physical structure in structural representation. It is computational structure that helps establish the second-order resemblance between internal vehicles and target domain. This move solves some of the problems with liberality that confronted structural representation. It introduces demanding constraints on what structures are candidates for representational status. Moreover, it accepts as candidates only those structures that are cognitively relevant, insofar as they play a computational role in enabling the computational capacities of the whole system under examination. Nevertheless, the liberality due to reliance on second-order resemblance is still present. Computational structures bear resemblance relations to many different target domains — Paris, Tianducheng, an indefinite number of mathematical operations, and so on. Modifying structural representation by placing computational structure mechanistically individuated at its core helps, but does not solve the problem of non-uniqueness of content. Representations still have as their contents many disparate things.

My proposal is to learn to live with this consequence by deflating the notion of representation. It might seem that denying robust reductionism about representation will hurl the theory in the throes of instrumentalism, or outright eliminativism. But this need not be so. Given the robust non-semantic account of concrete computation that the mechanistic view offers, it becomes possible to deflate the notion of representation while keeping to objectivism about, at least, representational vehicles, *i.e.* computational states and processes. This bring us back, once again, to interpretational semantics.

7.3 What about Interpretational Semantics?

The foregoing modifications to structural representation lead to a position according to which it is computational structure that carries much of the explanatory burden in an account of cognition. It is thanks to the fact that computational structures internal to the cognitive system bear resemblance relations to things in the world that

below.

the cognitive system is able to exploit them as models, as stand-ins. These structures allow the cognitive system to have "the structure of the world at [its] computational fingertips"[12]. The second-order resemblance relation between computational structures in the cognitive system, and things in the world is, at least in many cases, responsible for enabling appropriate complex behaviour.

The computational structure of internal states and processes lies at the basis of ascriptions of representational content. Content is to be seen as what explains the successful use of an internal state in the context of certain task domains, in which mechanistically individuated computational structure plays a central role. Cognitive systems are computing mechanisms, and (some) representations are those computational structures that, by instantiating the same relational structure of entities in the world, play a guiding role in complex behaviour.

Partially shifting the explanatory burden to computational structure is a move that, among the accounts on offer in the literature, only the mechanistic view of concrete computation allows, for the reasons we saw above. Such burden-shifting, in its turn, allows us to lift much of the weight traditional theories of cognition place on the notion of representational content. Representational content is not needed to individuate the relevant explanatory states, nor is it required to avoid triviality of concrete computation. This clears the way for a deflated notion of representation, one in which, as Egan (2010) puts it, content-ascription allows us to build a nexus between computational goings-ons, and the cognitive task at hand. This is in the spirit of interpretational semantics. Representations are invoked in order to reveal computations as cognitive processes — computations are interpreted, in light of the situation at hand, as simulating entities and processes in the world.

Given this deflated understanding of representation, the problem of non-uniqueness of content does not arise. As for interpretational semantics, the fact that a computational structure can be interpreted as modelling disparate target domains does not detract from its being explanatory for the cognitive task under investigation (however it is individuated). It is the suitability of the computational structure employed as a model of the target domain that provides us an explanation of why the organism behaved appropriately in a certain circumstance. Supplementing structural representation with the mechanistic view of computation leads to a form of interpretational semantics superior to the original versions of the theory advocated by Cummins (1989) and Ramsey (2007).

These claims will be further substantiated in the next chapter, in which I will flesh out with care the shapes my deflationary approach can take. This revamped interpretational semantics opens up two paths worth investigating: pragmatism about representational content, à la Egan; or a sort of 'mild reductionism', inspired by Dennett. I will explore each path in turn.

[12] Cummins (1996, p. 93.)

Chapter 8

Two Paths to Deflationism

It is time to take a look at how the type of deflationism about cognitive representation I propose can be cashed out in detail. A deflationary theory of representation on the premises that I have taken pains to lay down in the course of this work may take different forms. I will focus on two of them — pragmatism and mild reductionism — which strike me as more plausible, and more closely related to positions already found in the philosophical literature. There may be other deflationary paths to be explored, and even different ways in which pragmatism and mild reductionism may be cashed out. I will though limit myself to one version of each of the latter approaches.

Before proceeding, it is important to be clear on what is meant by 'deflationism' in the foregoing. The term has recently been used with a different meaning by Burge (2010), and thus, lest there be confusion, I want to contrast his use of the term with the one of most relevance here.

Burge (2010, pp. 293ff.) dubs 'deflationary views about representation' those theories that do not make representation a distinctively psychological notion — theories that thereby belong to the 'Deflationary Tradition'. Most theories of representation developed in the literature — and all that were examined in the present work — qualify as deflationary *sensu* Burge. For the notion of representation they end up with is such that it encompasses non-psychological phenomena: *e.g.* magnetotactic bacteria, thermometers, bug detectors in the frog tongue-snapping reflex, rudimentary maps in artificial and simple biological systems, circadian clocks in plants[1].

Burge sees this feature of theories in the Deflationary Tradition as deeply unsatisfying. The term 'deflationary' has for him a clearly negative connotation. Deflationary theories are too liberal about representation, detaching the notion from its explanatory role in cognitive science. This, Burge claims, seeds confusion, for it obscures the distinction between a liberal, non-psychological notion — which could be replaced by other terms, such as 'causal co-variation', 'structural resemblance', and so on — and a notion of representation that is proprietary to cognitive science, which is properly psychological[2]. Deflationary theories mislead by using a term that should apply only

[1] See also Morgan (2014).

[2] As Morgan (2014) points out, in a similar vein, both indicator representations and structural representation are in no way distinctively 'mental'.

to psychology to domains that have no relation to it. The notion of representation, Burge holds, is thereby overstretched, and applied to phenomena that are not genuinely representational, that do not call for explanation in representational terms.

According to Burge, moreover, the whole idea that representational content must be naturalised in order to secure the foundations of cognitive science is misguided[3]. In contrast, he argues that the crucial role played by the notion of representation in a successful science, *i.e.* cognitive science, which has produced fruitful theories and explanations, is enough to make the notion scientifically respectable. Representation is to be seen as a scientific primitive in cognitive science, an irreducible explanatory notion[4]. Burge's primitivism about representation in cognitive science is in clear opposition to the project I am pursuing. The tradition this work inscribes itself in is that of attempts to naturalise representational content, namely to explain content in naturalistically-acceptable terms.

I will not assess the merits and demerits of Burge's primitivism about representation. Suffice to say that representation seems to be the wrong candidate for an irreducible scientific notion. For one, representation may turn out to be an unneeded posit in explaining behaviour, if anti-representationalism should in the end prevail, and displace representationalism as the mainstream position in the cognitive sciences[5]. Moreover, by taking representation for granted, cognitive science is making its life easier than it should be: it is helping itself to a notion that carries a considerable explanatory burden, without worrying about whether that notion is naturalistic, or brings with it non-naturalistic commitments that may undermine its scientific credentials.

That said, it is important to keep in mind that my use of the term 'deflationary' applied to theories of representation, and in particular to the theory of representation that I am here developing, is in no way related to Burge's. What I mean by 'deflationary' and 'deflationism' has little if anything to do with the breadth of application of the notion of representation. Deflationism in my sense has to do with the following three features.

First, it concerns theories that see representation as carrying less of the explanatory burden in theories of cognition than is normally held, and thereby downsizing the metaphysical worries connected to the notion. If representation and content play a smaller, or at any rate, a different role in the explanation of cognitive capacities on the lines that I propose, their explanatory importance as well as their metaphysical import are deflated, and as a consequence at least some of the traditional problems with theories of content disappear — or so I argue. Let us call this aspect 'explanatory deflationism'.

Furthermore, the account of representation I put forward is deflationary inasmuch as it eschews the somewhat classical view, famously espoused by Fodor (1987), that the explanatory states in cognitive science are primarily individuated by their contents — a crucial component of what Egan (2014*b*) has dubbed 'Hyper representationalism',

[3]Rescorla (2013, p. 693) shares this view to some extent.

[4]Burge (2010, p. 308). Chomsky (1995) can also be interpreted along these lines.

[5]Ramsey's (2007) aim is exactly to argue that this is already the case in current cognitive science, appearances notwithstanding.

and Dennett (1991) has called 'industrial-strength Realism'. In contrast, I hold that cognitive states are not primarily individuated by their contents — the properties they possess that play a primary explanatory role in the cognitive sciences are not semantic properties, but computational ones. I will dub this aspect 'content deflationism'.

Finally, the view I propose is deflationary insofar as it rejects the requirement that a set of necessary and sufficient conditions for determining representational content be provided. This is akin to the first feature that Suárez (2004) includes in his 'deflationary strategy' toward scientific representation. As for Suárez (2004), my deflationary theory of cognitive representation involves individuating some general and non-exhaustive features that characterise representational content in cognitive systems — which I will call 'individuation deflationism'.

These three forms of deflationism are independent of each other. Being a deflationist about representation in one or two of these ways does not entail being one in the remaining sense(s). My account, nonetheless, embraces all three forms of deflationism[6].

One motivation for pursuing this deflationary path comes from the suspicion that the complexity and flexibility of the abilities sported by cognitive systems require that several different notions of representation and content be employed, on pain of leaving out of the explanatory purview some relevant cognitive phenomena. Views of representation closer to informational semantics may be relevant for explaining what feature detectors in sensory systems represent, and how they contribute to informing simple behaviour. They may also help explain the construction of more complex, structural representations that account for higher cognitive capacities, such as navigation, planning, and reasoning[7]. Pluralism about representation and content fixation is part and parcel of my deflationary approach.

Pluralism does not entail the rejection of necessary and sufficient conditions for representational content. It only entails that there may be several independent sets of sufficient conditions, each set defining a different type of representation. The disjunction of sets of sufficient conditions may be taken as a necessary condition for something to represent. The deflationary picture of representation I want to offer goes further. It sees representation as being strongly dependent on context; different contexts leading to different factors playing a role in the fixation of content. There is no list of relevant naturalistic factors for the fixation of content, no set of sets that includes all the sufficient conditions defining the plurality of processes that determine content.

Let us now move to the examination of what I take to be two main paths to which a deflationary view of representation leads. As long as terminological confusion with Burge's independent use of the term is avoided, I believe that 'deflationism' suitably describes the view of representation that I will now turn to explore in more detail.

[6]It is arguably also deflationist *sensu* Burge.

[7]Johnson-Laird's (1983) theory of mental models, based to a large extent on second-order resemblance, is an example of how explanations of complex, typically personal-level cognitive capacities in terms of structural representations may go.

8.1 A walk with the pragmatist

The first path, the one recommended by content pragmatists, is guided by the idea that the notion of representational content grabs a hold only in light of our explanatory interests. Representation and misrepresentation are notions invoked relative to specific explanatory purposes, and to measures of behavioural success dictated by the interests of theorists. The term 'representation' does not (necessarily) capture anything objective, out there in the world independently of our purposes and practices. Rather, it provides us with an explanatory handle adequately to describe the contribution of computational processes to the successful behaviour under investigation.

Frances Egan (1999, 2009, 2010, 2014*b*) has been a compelling advocate of such a view. Talk of representation in (computational) cognitive science, for her, is part of an 'explanatory gloss' that allows cognitive scientists to connect the computational goings-on with the cognitive task at hand. Representational content is an artefact of our investigation and theorising, not (necessarily) something really in the minds of cognitively complex organisms. Nevertheless, the view avoids eliminativism about content by insisting that representation plays a crucial role in cognitive science — an epistemic, rather than ontological one.

Let us briefly explore Egan's view, which inspires, but does not fully coincide with, the deflationary path that I want to put forward.

8.1.1 Egan's content pragmatism

Egan's analysis has as its starting point the role of the notion of representation in the cognitive sciences, rather than the more traditionally philosophical problems about representation that have exercised philosophers of mind. Her project consists in describing and making sense of the commitments and workings of our best sciences of the mind, with special attention to their apparently more promising (and mainstream) branch: computational cognitive science. According to Egan, computational explanations in cognitive science are function-theoretic in nature — they unveil which mathematical functions are computed by cognitive systems that help explain their behavioural success.

One of Egan's favourite examples is Marr's theory of how the early visual system detects edges, by calculating the rate of change in the intensity of light across the retina[8]. According to Marr, the early visual system sports this capacity because a part of it computes a mathematical function, the Laplacean convolved with the Gaussian. This mathematical function is general, inasmuch as inputs need not be light intensities and outputs edges, and is applicable to several other domains; it belongs moreover to a class of well-understood mathematical tools. Egan claims that this is the primary characterisation of cognitive mechanisms in computational cognitive science: what cognitive scientists look for in their explanations are mathematical, function-theoretic characterisations of the computations carried out by cognitive systems, or parts thereof.

[8]See Egan (2010, 2014*b*), Marr (1982).

The function-theoretic characterisation does not involve distal content. The system is explained, and its parts individuated, in terms of the algorithms they implement in computing the relevant explanatory mathematical function. Distal representational contents are not mentioned in the theory itself, the latter being characteristically 'environment-neutral'[9]. The computational cognitive mechanism that computes the Laplacean convolved with the Gaussian does so regardless of the environment it finds itself in, and therefore regardless of the causal connexions that input systems have with things in the world. Were the inputs to the mechanism transduced from sound waves rather than light intensities, the mechanism would still compute the same mathematical function over those inputs. These features are not restricted to this particular example, but are supposed to generalise to most, if not all endeavours in computational cognitive science[10].

Crucially, the relations between cognitive mechanisms and the environment, which typically are relied upon to determine representational content — according at least to mainstream theories of content, which are externalist — play no role in individuating the cognitive mechanism. Computational cognitive science individuates cognitive mechanisms not by their representational contents, but by their mathematical, function-theoretic, environment-neutral characterisation[11].

A place for representational content is nonetheless preserved, for appeal to content is unavoidable in explanations in cognitive science. The *explananda* of cognitive science are typically tasks that are pretheoretically characterised in distal terms, *i.e.* in terms of capacities sported by organisms in their interactions with their environments. Success in those tasks is assumed. In order to make clear how the mathematically-individuated mechanisms unveiled by computational cognitive science can be the *explanantia* of those distally-individuated *explananda*, an interpretation of the computational goings-on in terms of distal features of the environment is called for. That interpretation typically involves representational content: states and processes of the computational mechanism are seen as representing entities and processes in the world. In this way, a connexion is established between the computational mechanism, and the cognitive task it is meant to explain. Such interpretation is not part of the computational theory proper, which is environment-neutral, having no role for distal representational content. Rather, the representational interpretation of cognitive mechanisms is part of what Egan variously calls the 'explanatory' (2014*b*, 2014*a*), 'cognitive' (2010), or 'intentional gloss' (2014*b*). The gloss, by including reference to distal content, allows us to see how computing certain mathematical functions in a certain environment constitutes the performance of a cognitive task.

Beside playing the role of 'connective tissue' between computational story and pre-

[9] Egan (2014*b*, p. 122.)

[10] Egan (2010, 2014*b*) draws on several other examples from computational cognitive science, such as Shadmehr and Wise's (2005) theory of motor control.

[11] In her more recent work, Egan (2014*b*) argues that computational states have one kind of content essentially, *i.e.* mathematical content. I think this move introduces considerable and unnecessary complications, and should therefore be rejected *(cf.* Piccinini 2015, pp. 137-8). I will thus stick to formulations of the view that do not posit this new kind of content. See section 8.1.2.

theoretical *explanandum*, the gloss has also a heuristic value. It allows theorists more easily to grasp what the computational processes in the cognitive system are doing, and how the information flow inside the system proceeds. By seeing those processes as being carried out over states with distal representational content, the computational goings-on in the system become easier to understand, and synthetically to describe.

In sum, representational content, though not part of the purely function-theoretic characterisations that constitute the theory proper, plays a pragmatic role in allowing us to understand how cognitive computational systems work, and how their workings amount to the performance of pretheoretically characterised cognitive tasks. Content is fixed not only by means of natural relations between certain cognitive states, and processes and entities in the world, but also by a host of pragmatic considerations motivated by the explanatory aims and practices of theorists[12].

Ascribed contents must be salient and tractable, and should not be unwieldly disjunctive or opaque. The representational interpretation may not perfectly match the computational description of the system — one-to-one mappings between computational states and contents may be given up — provided that such idealisation increases expository clarity. Moreover, depending on the level of generality that is supposed to be achieved by the cognitive gloss, more or less general contents may be ascribed. The same computational system may contribute to different cognitive tasks, involving different types of input and output (*e.g.* visual *vs.* auditory), in different contexts. Depending on explanatory aims and interests, content ascription may try to capture the specific contributions made in the different contexts, ascribing visual contents in one type of case, and auditory contents in another; or it may strive for greater generality by ascribing disjunctive contents, *e.g.* both visual and auditory; or it may strive for yet greater generality by ascribing more proximal representational contents, *e.g.* changes in intensity of the signal.

Misrepresentation is possible in content pragmatism, though it grabs a hold only relative to a choice of ascribed content[13]. It is only given a certain pragmatic background in which a choice has been made of what content to ascribe to a certain computational state that questions about representational correctness and misrepresentation can be evaluated.

Take the (in)famous example of frogs' tongue-snapping reflex. In some explanatory contexts, *e.g.* when the behaviour of interest is the ability of the frog to feed itself, internal states of the frog may be ascribed the content *that there is food at x,y*. Every instance in which the reflex behaviour is produced in absence of food, but, say, in presence of dark moving pellets of inedible material, will count as instances of misrepresentation. In a different explanatory context, *e.g.* when the explanatory aim is that of understanding how the visual system of the frog works, pragmatic considerations may lead to ascribing to internal states the content *that there is a small dark moving spot at x,y*[14]. Those instances that counted as misrepresentations in the context of

[12] Egan (2014*b*, pp. 125ff.)
[13] Egan (2014*b*, p. 127.)
[14] Egan (2015).

explaining frogs' ability to feed themselves are, in this context, instances of correct representation instead. What remains unchanged with the change in explanatory context is the mathematical function(s) that the reflex-generating mechanisms in the frog compute (whatever they are). The function-theoretic description remains constant, while the computational states and processes have different contents ascribed to them in the different cognitive glosses.

This latter point further clarifies in which way the computational states and processes are not individuated by the contents they bear. Individuation in computational cognitive science involves primarily the function-theoretic characterisation, which remains constant with change of explanatory context, embedding environment, different weight given to different pragmatic considerations, and so forth. Representational content, on the other hand, may vary with changes in each of the latter factors. These different content ascriptions do not affect the individuation of the computational states and processes that are the *explanantia* of computational cognitive science. What is essential to those *explanantia* is the mathematical function they compute, not the representational contents that may be ascribed to them.

Egan's approach is in the spirit of the deflationism that I advocate. Indeed, she has recently labelled her view 'deflationary'[15]. As she puts it, her deflationism is objectivist about the explanatory vehicles in cognitive science, *i.e.* computational states and processes, whereas it is pragmatist for what regards representational content[16]. As it stands, I believe that Egan's view has trouble doing justice to the former claim. Given her allegiance to an unsatisfactory view of concrete computation, she has trouble securing a non-trivial, observer-independent notion of concrete computation, and thereby of computational vehicles. The mechanistic view of concrete computation that I have presented and developed in Part II can providentially come to the rescue.

8.1.2 Content pragmatism and concrete computation

Egan seems to[17] be working with a notion of concrete computation akin to that put forward by Cummins (1989)[18]. She seems to subscribe to the view that computation involves a mapping between steps of an algorithm, and states of a physical system — computational descriptions providing 'a formal characterisation of a device'[19]. It is not clear whether, at least in her early work, Egan rests contented with a simple mapping view of concrete computation with the addition of pragmatic constraints, as

[15] Egan (2015).

[16] More precisely, Egan (2015) claims to be an objectivist about representational vehicles, and a pragmatist for what regards representational content. I find this phrasing puzzling, for it is not clear how there can be objective representational vehicles part of the computational theory proper, whose content is though observer-dependent and not part of the computational theory. To be fair, the vehicles can count as representational, for Egan, in light of the mathematical content they have, but not of their cognitive content. Since I reject Egan's more recent semantic view of computation based on mathematical content, I will phrase the claim in terms of computational vehicles, rather than representational ones. Ramsey (2015) presses a similar point, though not directed at Egan.

[17] I use 'seems to' throughout this paragraph given that it is not always transparent what Egan's commitments are regarding concrete computation.

[18] See Egan (1999, 2010).

[19] Egan (1999, p. 181.)

examined in section §3.2; or whether she also wants to include causal constraints. As argued in section §3.1, endorsing the simple mapping view leads to triviality of concrete computation, and unlimited pancomputationalism. Adding pragmatic constraints on computational individuation would jeopardise Egan's commitment to objectivism about computational vehicles. More recently, Egan (2014*b*, p. 116) has more openly appealed to causal constraints on the mappings relevant for determining computations. Since causal mapping views are stronger and more plausible than simple mapping views, I will stick to the more charitable reading in this section, and take Egan to endorse a causal mapping view on the lines of Cummins'[20].

Though non-semantic, computations are typically semantically-interpretable, according to Egan. Concrete computation involves two mapping functions: f_{R} and f_{I}. The former is the realisation function, which maps computational states and processes onto physical states and (causal) processes in the physical system. The latter is the interpretation function, which maps computational states and processes onto contents. Egan's views have changed with regards to f_{I} in the course of the years. In her (1999), what is crucial is f_{R}, while f_{I} is not essential to concrete computation, computation being non-semantic. This view, indistinguishable from Cummins', brings with it the problems presented in section §6.4. I will not rehearse those arguments again here. Suffice to point out that this view, unless constrained by further factors, leads (at least) to limited pancomputationalism, thereby failing appropriately to capture the domain of physical systems pretheoretically considered to be computational.

Furthermore, though it preserves objectivism about computation, this view makes concrete computation into a trivial matter, and the existence of computational vehicles ubiquitous and uninteresting. For Egan's objectivism about computational vehicles to be a less trivial claim, pragmatic constraints would likely have to be appealed to in order to select, out of all the computational physical systems, the ones that can be fruitfully computationally explained. And, out of the latter, which of the computations they perform are explanatory of their behaviour. In the case of cognitive science, pragmatic considerations would be invoked twice: in determining which concrete computations are explanatory, and in determining the representational content ascribed to those computations, which appears in the explanatory gloss.

I think that such a view is coherent, and perhaps defensible. But it leaves ample margin for improvement. Its treatment of concrete computation can be bettered by having recourse to the robust mechanistic view defended in Part II, which is non-trivial and does not entail pancomputationalism. Embracing the mechanistic view avoids the need for pragmatic considerations in determining which concrete computations are explanatory. As we have seen, the mechanistic view narrows down considerably the domain of physical systems that compute, and provides the tools to individuate the computational structures of computational mechanisms.

More recently, Egan (2010, 2014*b*) has defended a heterodox semantic view of compu-

[20]Her indebtedness to Cummins' view of computation is clear in her (2014*b*), as well as in her early work. In her (1999) she refers to Cummins (1989) as one of the theorists providing the more accepted theory of concrete computation.

tation. According to her new view, both $f_{\mathrm{R}}$and f_{I} are essential to concrete computation. However, in contrast to traditional semantic theories, f_{I} does not map computational states onto distal representational contents, but rather onto 'mathematical contents', *i.e.* abstract mathematical objects such as numbers, and mathematical operations. Computational states essentially represent, though what they represent are numbers, mathematical operations, arguments and values of mathematical functions. A further, cognitive interpretation, maps mathematical contents onto representational contents (which she dubs 'cognitive contents'). This latter interpretation function is though not part of the computational theory, but rather of the explanatory gloss. So pragmatism about representational, or cognitive, content is still in force.

I take this move toward a semantic theory of concrete computation, albeit of a non-traditional kind, to be misguided. The difficulties pointed out in section 3.3.2 that semantic theories have to face apply, with the added worry of having to make sense of the notion of mathematical content itself. Which kind of relation must there be between physical states and processes for them to represent abstract mathematical entities? Egan does not offer an account, though she sees the prospects of coming up with a naturalistic explanation of that representational relation to be dim. The lack of a clear account of how computational states get their mathematical content is problematic, since as Piccinini (2015, p. 138) points out, computational states and processes are compatible with different assignments of mathematical content[21]. Mathematical content would thereby be non-unique, and since mathematical content helps individuate concrete computations, the upshot is a semantic version of the multiplicity of computations problem. The same physical goings-on can have different mathematical contents, and thus implement several different concrete computations simultaneously.

Egan's more recent view of concrete computation is at best crucially incomplete and, as it stands, deeply problematic. The same can be said of her previous view. In both flavours, Egan's theory is ill-equipped to do justice to one of its own aims: being objectivist, in a explanatorily helpful way, about computational vehicles.

The mechanistic view, for the reasons already rehearsed at length, is up to the task of securing what Egan's theory requires. This is one of the main contributions that the foregoing work makes to the pragmatist version of deflationism about content: it provides a robust notion of concrete computation much more adequate in ensuring that some physical systems have states and processes that are computational in an objective, non-trivial way. Since computational vehicles, in the content pragmatist picture, are supposed to carry most of the explanatory burden in cognitive science — representation being part of a helpful gloss — it is crucial that they be objective and non-trivial. Otherwise, their explanatory role and theoretical importance would be jeopardised, and the content pragmatist would be in dire straits.

The deflationary path recommended by content pragmatism is made more cogent by incorporating the mechanistic view of concrete computation in the account. Let us now look at how such an amended content pragmatism looks like.

[21]See also Rescorla (2013, pp. 687-8.)

8.1.3 Content pragmatism amended

I will refer to this pragmatic, deflationary theory of content based on interpretational semantics as 'pragIS'. The interpretation function that maps computational states and processes onto representational contents is not purely mathematical, as for Cummins (1989), but is determined by pragmatic considerations. The relevant interpretation function, the relevant mapping, is the one that allows explaining the success of the organism in the task at hand in an explanatorily illuminating and interesting way. As we have seen, Egan (2014*b*) dubs it 'cognitive interpretation'. PragIS is not a naturalistic theory of content, for it does not explain representation in naturalistic, non-intentional terms. The notion of interpretation at play is itself intentional — it depends on the purposes and interests of cognitive scientists. This fact is not particularly problematic as long as we keep the ambitions of the theory suitably constrained. Recall that we are here concerned with a theory of representational content, and not of intentional content, as defined in Part I. Accounting for these two different notions of representation and content may well be largely independent projects, involving quite different factors and considerations.

The pragmatist flavour of IS makes explanation of representational content parasitic on intentional content. The former depends on the latter. However, this is no strong objection to the foregoing — it may well be that there is no naturalistic, objectivist account of representational content. Whether there is one for intentional content is another, plausibly much more complicated, question. The content pragmatist need not take a side on this further debate, though non-robustly reductionist options are available [22].

The pragmatist-flavoured interpretational semantics I propose is objectivist about concrete computation and computational structures, courtesy of the mechanistic view of concrete computation, but it is deflationist for what regards representational content. The view does not invoke mind-dependent notions to account for computational structure, but it does invoke them for what regards ascription of content. The foregoing account is able to make space for the possibility of misrepresentation. Misrepresentation takes place when the computational structures driving a specific behaviour fail to lead to successful behaviour in the task at hand. Misrepresentation may be due to interference, malfunction, inadequate circumstance, etc. What the task is, and what counts as success depend on pragmatic considerations — the explanatory interests of scientists.

We start with behavioural success in a task, and look for an explanation for it. Success, as Ramsey (2007) and Egan (2014*b*) point out, is presupposed — it is the explanatory target we start with, and is thus dependent on pragmatic considerations. Representational success is, as it were, an explanatory consequence of behavioural success (of a certain kind). The deflated notion of representation at work in pragIS involves an inversion of sorts of the relations between representation and successful behaviour. Ap-

[22] For instance, (Dennett 1971, 1987*a*). Though Dennett's position, especially in his early work, was close to instrumentalism or pragmatism, later versions of his view are open to a more naturalistic reading. I will explore such a reading in section §8.2.

propriate representations explain behavioural success; but behavioural success is what makes the representations appropriate. Computational structures explain how come the organism behaved as it did, and content-ascriptions connect those structures to things in the world, making clear why those structures enabled the behavioural feat. Behavioural and representational success go hand in hand in this explanatory strategy, and so Burge's (2010) point about the independence of representational and behavioural success does not apply. It concerns only attempts at naturalising content, for which the need to distinguish representational and behavioural success is required so as to avoid the putatively wrong-headed reduction of the former to the latter. However, the pragmatist flavour of interpretational semantics is not a naturalistic theory of content, and therefore need not worry about picking an inadequate reduction basis.

PragIS puts into question the assumption that only accounts of representational content that are naturalistic are on the right track. Rather, it claims, the need for naturalisation of representation — at least of the subpersonal kind — disappears once we take a different approach to what representation is, and what we require the notion to achieve in our scientific theories. The approach pragIS recommends partly turns the metaphysical problem of content into an epistemological one. Ascription of content to computational structures is based on the fact that those structures are responsible, given the resemblance relations in which they stand with things in the world, for the successful behavioural outcome. A large part of the burden of explanation is on computational structures and their causal powers, representation being a way to connect computational goings-on with the cognitive task to be explained.

Appeal to representational content plays an important explanatory role, inasmuch as it equips us with a way of grasping the relevance of computational processes in making successful behaviour possible. But ascription of representational content fundamentally depends on the interpretation given by theorists of what the task at hand is, and of how the computational structures allow it to be successfully carried out. There is no question of internal states essentially bearing semantic properties. In the foregoing picture, what is essential to the relevant states is their computational structures.

Cognitive representation is nothing mysterious, nor in need of metaphysical vindication. And wild non-uniqueness of content does not pose a threat. The deflated notion of representational content at hand is not harmed by the degree of liberality of second-order resemblance that remains, for it is unproblematic that the same computational structure can bring forth successful behaviour toward many different target domains. Liberality is not a problem for the explanatory project of which the notion of representation is an important part — all that should concern the theorist is that those structures are playing a role in successful behaviour toward the target domain at hand, regardless of their potential role in other behavioural domains. Once those structures are individuated, and their role in enabling successful behaviour is established, representational content can be ascribed to the elements in the structures as a way to reveal how those elements, and the computational processes they partake in, are relevant to the behaviour under investigation. The fact that the same computational structures

could equally lead to successful behaviour toward different target domains is irrelevant for the explanatory purposes at hand, and will thus not call for additional contents to be ascribed to the structures.

In sum, representational content plays an explanatory role in the pragmatist flavour of IS, albeit a secondary one — that of an explanatory or cognitive gloss over what is doing the work, *i.e.* computational structures mechanistically individuated. The fact that the same computational structures may be effectively employed to solve different tasks is unsurprising. Moreover, it does not detract from its explanatory role in explaining specific behaviours. Second-order resemblance comes in to help account for why on certain occasions computational structures lead to successful behaviour, whilst in others they do not. In the former case, there is a second-order resemblance between the computational structure and the target domain, while in the latter there is not, or not enough.

Content pragmatism has the tools to avoid indeterminacy of content. First, indeterminacy of content is a problem only if there is a commitment to there being cognitive states that essentially bear content. The content pragmatist can deny that representational content is an objective property of cognitive states, and reject the view that cognitive states are essentially individuated by their contents. Second, representational content, being part of the explanatory gloss, is ascribed with pragmatic considerations in view. Depending on explanatory aims and interests, different specific contents may be ascribed to the same cognitive state. Determinacy is achieved if and to the extent that it is explanatorily useful relative to the relevant pragmatic considerations. Misrepresentation is similarly handled. It is only given a certain pragmatic background in which a choice has been made on what content to ascribe to a certain computational state that questions about representational correctness or misrepresentation can be evaluated.

A further source of liberality that was problematic for theories of representational content based on second-order resemblance — namely, that we should allow approximate resemblances to count, as perfect resemblance is too demanding a requirement — is innocuous on pragIS. The second-order resemblance is good enough to the extent in which it enables the behaviour under investigation to be successful. The elements of the relevant structures have determinate contents, for the connexion of relevance is that between the computational structures, and the target domain. There is no sense in ascribing slightly inaccurate contents just because the pertinent second-order resemblance is not perfect. Computational structures drive appropriate behaviour because the resemblance is good enough, and that is good enough for determinate content-ascription.

8.1.4 Content pragmatism assessed

In recent years, partly due to the frustration felt in some philosophical quarters with the project of naturalising content, and partly thanks to Egan's compelling advocacy, content pragmatism has attracted considerable attention, as well as criticism. The most vocal direct opposition has been put up by Ramsey (2015), Neander (2015), and Bechtel (2016). Ramsey and Neander make use of a similar strategy in arguing against

content pragmatism. The view is supposed to provide an alternative path in the philosophical attempt of understanding representation and content, eschewing reductionism, eliminativism, primitivism (and dualism) — the traditional positions, often taken to be exhaustive of the conceptual landscape[23]. Neander and Ramsey argue that this is an illusion: content pragmatism does not offer a coherent position, and thus ends up collapsing into one of the traditional views.

Bechtel, on the other hand, focuses on content pragmatism's putative inability to capture the way cognitive scientists use the notion of representation in their theories. I will start with Bechtel's critique, and then move on to Neander's and Ramsey's.

Bechtel's critique

A first worry concerning pragmatism about content in general, and therefore also pragIS, is that it is radically revisionary of the practices of cognitive scientists. It would seem that cognitive scientists make use of representations not as explanatory glosses to be provided once their theories are complete, but rather as hypotheses about states and processes that objectively exist in cognitive systems. Bechtel (2016) examines one case study, the discovery and development of theories of spatial representation in the rodent brain — which came to be known as 'cognitive maps' (Tolman 1948, O'Keefe & Nadel 1978). In the decades that followed the original discovery of place cells, grid cells, head-direction cells, and other processes that play a role in spatial representation and navigation have been found[24]. Bechtel's aim is to show, *contra* Egan, that the positing of representations, rather than working as a mere gloss, actually involves strong ontological commitments, drives new discoveries, and deeply informs theorising in cognitive science.

> ... treating brain processes as representations is foundational to these research projects as they have been pursued. The research pursuits are focused on identifying what neural processes serve as representational vehicles and especially determining what they represent. The research efforts are designed to answer these questions and it would be difficult to understand why the researchers pursue these projects if their goal was not to identify representations and determine what they represent. (Bechtel 2016, p. 1289)

Bechtel describes in great detail the process of discovery and theoretical development that have marked the work on place cells, grid cells and head-direction cells for the past four decades. He uses this careful reconstruction of scientific progress in this area, what motivated it and what it was directed at explaining, to argue that a commitment to the objectivity of representations was essential for the scientific project to go on. The whole process would have seemed absurd and meaningless were talk of representation to be taken as merely providing an explanatory gloss, rather than as capturing objective states and processes in the rodent cognitive system. Though Bechtel's careful treatment of this case study is historically and philosophically interesting, I think that he fails in his

[23] See section §1.3.
[24] Moser et al. (2008).

attempt to, by means of this case study, undermine pragmatism about representational content.

Bechtel is probably right that *prima facie* content pragmatism is revisionary. As he correctly points out, most of cognitive science traffics in representational talk. However, more must be shown if his line of objection is to be successful: he must establish that talk of representation is justified and substantive, and not a mere matter of scientific heritage. Ramsey (2007) investigates this question, and concludes with a largely negative answer: according to him, most of contemporary cognitive science employs a notion of representation based on mere causal correlation, what he dubs the 'receptor notion'[25]. The receptor notion of representation, he then argues, fails to distinguish representations from mere causal relays. Most talk of representation in the cognitive sciences could be replaced with talk of causal relays, or mediating causal states, without loss of scientific value[26]. Appeal to receptor-representations, for Ramsey, does not buy us anything explanatorily useful in comparison to appeal to causally mediating states, and should thereby be excluded from theorising in cognitive science. Ramsey takes S-representation to be the only robust notion of representation on offer.

Ramsey's sceptical results about the presence of a substantial notion of representation in contemporary cognitive science have been compellingly countered by Sprevak (2011) and Shagrir (2012*c*), at least for what regards some domains of enquiry. At any rate, the cautious note that Ramsey urges on us remains in force: talk of representation in the cognitive sciences, though widespread, may be largely empty, and we must examine in each case whether it is playing its proper explanatory role. The fact that cognitive scientists routinely deploy the term 'representation' should not lure us into believing that they are actually positing the objective existence of robust representations in cognitive systems — they might be simply talking about causal relays.

Most importantly, Bechtel's case study is fully compatible with the tenets of content pragmatism. As the content pragmatist points out, the notion of representation plays a crucial role in cognitive science inasmuch as it allows us to connect the computational *explanantia* with the distally-individuated *explananda* of cognitive science. It is natural for cognitive scientists to start with the cognitive task they want to explain, which is individuated in terms of the relationships between organism and environment — in the case at hand, the ability of rodents to navigate space. That choice of explanatory target plays a central role in directing scientific research toward the discovery of the computational mechanisms behind the feat.

The fact that an 'early and integral step' in cognitive science is that of "using in-

[25]This is indeed how Sullivan (2010) defines the notion of representation at play in cognitive neuroscience and neurobiology. She then goes on to argue that in many cases talk of representation only plays a minimal, heuristic role in research and explanation.

[26]Markman & Dietrich (2000) hold that causal mediation is the core of the notion of representation. Their account is thereby extremely liberal, and fails to explain what is special about appeal to representation in the cognitive sciences, as opposed to other sciences in which causal mediation also plays an explanatory role, but there is no talk of representations. By Ramsey's lights, this view does not capture what cognitive representations are supposed to be in our scientific explanations, therefore misusing the term as much as those that subscribe to the receptor notion of representation.

vestigations of content to help identify the vehicles"[27] is in harmony with content pragmatism. Identifying behavioural success in an distally-individuated cognitive task is typically the first step in research in cognitive science: it is the choice of *explanandum* that then constrains the scientific investigation that follows. This does not mean that talk of representation must be taken as ontologically committing; it may well be, as the content pragmatist urges, a useful way of connecting goings-on in the cognitive system with the behaviour to be explained: an explanatory gloss.

After determining the explanatory target, cognitive scientists go on and try to figure out what internal computational processes can explain behavioural success, what the computational vehicles are, *e.g.* in spatial navigation. This involves figuring out which stimuli neurons respond to, as well as their causal relationships with other neurons and networks. This allows cognitive scientists to individuate the neurons and networks that are good candidates for being the computational mechanisms behind the successful performance of the cognitive task. Once plausible candidates are found, it is only by taking them to be representing (or, more weakly, carrying information about) entities and processes in the environment that the connexion between the computational mechanisms, and the cognitive task target of the explanation can be established.

The fact, repeatedly appealed to by Bechtel (2016), that cognitive scientists are interested in finding out what neural states represent is perfectly in line with what the content pragmatist holds. It is a crucial step in trying to individuate the computational mechanisms that explain successful cognitive behaviour. The content pragmatist does not deny the importance of the notion of representation in the 'context of discovery'[28]. The gloss that representation provides not only allows keeping the explanatory target always in view, but it also makes simpler for scientists to understand the role that computational goings-on in the cognitive system play in bringing forth successful completion of the task. However, this does not entail that in the 'context of justification' an ontological commitment to cognitive representations is necessary. It does not follow from anything that Bechtel says about how research in rodent navigation has proceeded that such an objectivist position about representation needs to be taken to do justice to scientific practice. That practice is compatible with the tenets of content pragmatism as well. Therefore, I submit, the content pragmatist escapes Bechtel's objection.

Pragmatism and primitivism

The second line of attack against content pragmatism involves claiming that the view is too unstable, its instability making it collapse onto one of the traditional views of representational content. Neander (2015) worries that content pragmatism might fall into some form of dualism, or rather primitivism[29]. For the content pragmatist, the

[27] Bechtel (2016, p. 1291.)

[28] I am using the term, and its usual companion, in a somewhat loose sense.

[29] Neander, as well as Egan (2015), conflate dualism and primitivism, which strikes me as odd. Primitivism in the lines of Burge (2010) can be compatible with naturalism — it takes representation to be a scientific posit in no need of naturalisation, as much as fundamental entities posited in physics do not call for naturalisation insofar as they are part of successful and fruitful sciences. This view need not lead to dualism.

argument runs, content-ascription depends on explanatory aims and interests which are ultimately based on the intentions of organisms (humans). However, intentions are intentional states with content. For the account to work in a non-circular way, the content of the intentional states grounding the aims and interests of humans seems to have to be taken as primitive.

Egan insists that her account is targeted only at representational content (short of beliefs and desires), and not at the intentional content of propositional attitudes. She can hence avoid falling into primitivism, leaving open the question of the nature of intentional content. The content of cognitive states short of beliefs and desires is parasitic on intentional content, for it is the latter that helps determine the content of the former. But then, Neander presses on, content pragmatism accounts for only a part of the problem of content, being completely silent on intentional content, on which representational content depends. The content pragmatist would thus fail to give a full account of how content in general is possible in our world. Moreover, they would have to subscribe to a heterodox view of how the naturalisation project should go. The strategy typically considered to be more promising has been to try and give a naturalistic account of simpler cognitive states, short of beliefs and desires, to then later build the account up, perhaps by adding extra factors, to also encompass more complex, intentional states, such as propositional attitudes.

These are fair worries, I believe, but they fall short from undermining content pragmatism, let alone make it collapse into dualism or primitivism, or even perhaps robust reductionism about intentional content. The view does not collapse into dualism or primitivism because the kind of content it appeals to in helping to determine representational content is intentional content. Being a different kind of content, intentional content is plausibly to be explained by a different theory, targeted to that domain of phenomena[30]. Moreover, the content pragmatist need not endorse robust reductionism about intentional content. As I have already hinted, non-robustly reductionist views of intentional content are available. Dennett's (1981, 1987*a*) Intentional Stance view is supposed to account for the intentional content of propositional attitudes, without offering a straightforward reduction of the latter[31]. The game is still open as to which kind of approach to intentional content will eventually prevail[32].

Pragmatism and eliminativism

Content pragmatism and content eliminativism share one crucial element: both views deny objectivity to representational content. However, content pragmatism and content eliminativism are at odds in another crucial point: while the latter advocates elimination of the notion of representation from cognitive science as lacking explanatory value, the former claims that cognitive science cannot do without it. According to the content

[30]Cummins (1989, pp. 12-13, 88.)

[31]In a reply to Neander (2015) during the MindsOnline 2015 Conference, Egan seemed attracted to a view on these lines. A related account, also compatible with this line of argument, is the measure-theoretic view of propositional attitudes put forward by Matthews (2011).

[32]See Haugeland (1990) for a only partially out-dated overview of the options.

pragmatist, the explanatory gloss that accompanies computational explanation in cognitive science cannot be eliminated. For that gloss in terms of representational content plays an essential role in allowing cognitive science to explain what it sets out to explain: distally-individuated successful behaviour.

As Neander (2015) points out, content pragmatism would collapse into eliminativism were the pragmatist to hold that future advancements in the cognitive sciences may discharge the need for the gloss. In that case, the explanatory gloss would be a non-essential epistemic device that we now need because of the early stage of development of the cognitive sciences. In the future, with a fuller picture, and a better understanding of how it all works computationally and neurally, the explanatory gloss, with all its representational baggage, would be done away with. If content pragmatism amounted to that view, then it would indeed be a species of eliminativism.

But on the contrary, content pragmatism sees the explanatory gloss as essential to cognitive science, something that further scientific developments cannot eliminate. Though representational content is not part of the computational theory proper, it is an essential part of the explanatory gloss, without which it would be impossible to connect the *explanantia* and *explananda* of cognitive science. Therefore, content pragmatism does not give way to eliminativism.

Pragmatism and robust reductionism

The most insidious challenge to content pragmatism is avoiding its collapse into some form of robust reductionism about representational content. I will examine two arguments that purport to show that content pragmatism, in some way or another, fails to provide a principled alternative to robust reductionism about content. Since it is the objectivism that marks robust reductionism that underlies the crucial points in the discussion, I will mostly focus on the former. I will try and defuse these arguments, securing the coherence and interest of content pragmatism as a genuine competitor in the search for a satisfying theory of the nature and role of representational content.

Ramsey against the argument from environmental neutrality An argument that plays an important role for Egan in motivating content pragmatism is the argument from environmental neutrality of computational explanation. Egan (2009, 2014*b*) shows that computational mechanisms that contribute to a cognitive task in one type of organism in one type of environment (say, computing shape from shading) can be transferred to a different environment or a different organism and play a different role (say in audition), despite the fact that they still compute the same mathematical function. Thus different representational contents are ascribed to one and the same computational mechanism when embedded in different organisms and environments. This fact, Egan believes, argues against objectivism about content — it would show that representational contents are not essential to computational mechanisms, since ascription of the the former varies wildly when the latter are kept fixed, and only the embedding context is changed.

Ramsey (2015) has recently argued that this argument will not do. I believe that he is right — as it stands, Egan's argument does not provide reasons to prefer content pragmatism over objectivism. The fact that the same computational mechanism, by computing the same mathematical function, can have different contents when embedded in different organisms and different environments does not jeopardise its having fairly determinate contents in the actual organisms it is embedded in. Computational mechanisms have been plausibly selected for the functions they compute because they contributed to successful behaviour of members of a species in one or more types of environment — alternative contents they might have had given different evolutionary stories or different embedding contexts are irrelevant.

To avoid the objection, the content pragmatist must be more radical in their claims. They must claim that representational content is ascribed to different computational states and processes, and different combinations thereof, given different contexts and cognitive tasks in the same type of organism, and the same types of environment; and this due to heuristic and pragmatic reasons tied to our interests and practices[33]. If this is so, then content is not essential to computational states even in the same organism in the same environment — an outcome difficult to square with robust reductionism about content, given the considerable indeterminacy of content and of vehicle individuation across contexts that follows[34]. This is largely an empirical hypothesis, which only empirical work can vindicate or prove wrong[35]. It is also a warning: unproblematically endorsing the robust reductionist view of representation and vehicle individuation can skew theorists' and scientists' interpretation of the available data, generating puzzles and questions that would be misplaced given a more flexible view of the role of representational content in cognitive science.

Glossing reality The final argument against content pragmatism that I will examine is, I believe, also the most threatening. It tackles heads-on a crucial difficulty that content pragmatism has to face, namely providing grounds for seeing representational content as explanatory, whilst withholding objectivity to the notion. The content pragmatist must answer the pressing question: if talk of representational content cannot be eliminated from cognitive science, even after future progress, why then reject the idea that content ascription is capturing something objective about cognitive systems?

Critics of content pragmatism argue that it cannot justify the fracture between explanation and ontology that lies at the heart of the view[36]. Importantly, this fracture is not a general one; it does not apply to every, or even most, scientific posits —

[33]See the example of pragmatist treatment of the frog's tongue-snapping behaviour in section 8.1.1.

[34]This is not incompatible with content objectivism *per se,* but it is incompatible with mainstream robust reductionism about representation, which sees representations as stable and repeatable cognitive structures. Non-robust reductionist views are in the cards, and I will put forward one such view in section §8.2. At any rate, the pragmatist's insistence that such variety of content ascriptions stems at least partly from heuristic and pragmatic considerations is incompatible with objectivism *tout court.*

[35]Such a hypothesis is in line with neural reuse theories. For a review of the theories and the evidence for them, see Anderson (2010). See also section §8.4 below.

[36]I thank an anonymous reviewer to *Topoi* for framing the criticism at hand in this clear and pointed way.

a position characteristic of full-blown scientific pragmatism. By content pragmatism lights', the fracture comes in for representational content, but not for other scientific posits, such as computational vehicles and processes. The guiding idea is to cling to objectivism about the latter, while shifting to pragmatism about the former. Content pragmatists must provide principled reasons to hold that a notion that is explanatorily ineliminable from cognitive science — representational content — should not be seen as ontologically committing; whilst other putative explanatorily ineliminable notions — such as computational mechanisms — should be so seen. Unless such principled reasons are provided, the critic presses on, content pragmatism is unjustified. Why set the pragmatism/objectivism border there, rather than somewhere else, or not at all?

I think there are good, albeit defeasible motivations for placing the border right there. Cognitive science aims at furnishing a fully naturalistic account of what cognition is, and how it works. Notions whose naturalistic credentials are dubious should not compose the theory proper, on pain of endangering the scientific status of the field. Representation and content are clearly problematic on this regard. It is both due to their apparent resistance to naturalisation, and to their explanatory role in cognitive science, that philosophers have been so keen on trying and giving robust naturalistic treatments of those notions.

It is an open question whether representation and content will ever be naturalised, at least in the way robust reductionism would want. This provides some justification for a guarded scepticism toward an objectivist take on those notions. It also provides some justification for quarantining representational content in a pragmatically-motivated, non-naturalistic explanatory gloss. Perhaps future work will give us the much sought naturalisation of representational content[37]. If this should be so, content pragmatism would lose most of its motivation. However, until that happy day arrives, if it does, content pragmatism is justified in its project of pursuing a different path — an alternative way of seeing the role played by representational content in the cognitive sciences, *i.e.* not as part of the naturalistic theory proper, but as part of a supplementary non-naturalistic explanatory edifice, built by us, for us, in light of our interests, capacities, and aims.

This is just the first step of the rejoinder. I have so far showed only that, given our current knowledge, representational content possibly lies outside the border of what we should take to be objective features of the world. Similar considerations suggest that, as content pragmatism would want, computational mechanisms are to be found on the inner side of that border. For concrete computation, *i.e.* computation in physical systems, seems a more promising candidate for naturalisation than representational content. Despite some early scepticism about the prospects of naturalising computation (*e.g.* Putnam 1988, Searle 1992), recent proposals are much more robust, in particular the mechanistic view (Piccinini 2015, Milkowski 2013, Fresco 2014), as we have seen in Part II. If concrete computation is a natural phenomenon, there is no impediment to its appearing in scientific theories. The fate of content pragmatism thus hinges on whether

[37] And perhaps even the present work. See section §8.2.

concrete computation, in contrast to representational content, will be satisfyingly naturalised.

In sum, there are plausible, though defeasible reasons to believe that computational states and processes are acceptable components of a naturalistic, objectivist story about cognition. And there are equally plausible, and equally defeasible, reasons to believe that representational contents are not. This is justification enough to hold, at the current state of play, that the localised fracture between explanation and ontology may fall where content pragmatism claims it does. Though this may prove to be wrong given future developments, it is not an incoherent or ungrounded position.

Even conceding that there are cogent reasons to uphold the localised fracture between explanation and ontology that content pragmatism advocates, the critic may still not be satisfied, and rightly so. For now content pragmatism seems to flirt dangerously with eliminativism. The burning question thus becomes: if representational content is a bad candidate for figuring in a naturalistic theory of cognition, why not get rid of the notion entirely, as the eliminativist recommends?

To assuage that worry, the content pragmatist reminds that representational content can never be eliminated from the explanatory gloss because of the nature of the *explananda* in cognitive science. Since those *explananda* are characterised as cognitive abilities having to do with robust successful interaction between organism and environment, the only way to make the computational *explanantia* cogent as explanations of those abilities is to see their components and processes as representations of the body and environment. A commitment to objectivism about content does not follow, but neither does an elimination of the notion from explanation in the cognitive sciences.

Content becomes explanatorily ineliminable once we see physical systems as cognitive systems, bringing thus to bear our explanatory interests in making sense of their behaviour, understood in their turn as cognitive abilities. It is only when we take a specific perspective — a 'stance' — toward physical systems that representational content becomes something we cannot do without in making perspicuous how the computational goings-on explain behaviours characterised as cognitive. Taking such a stance brings with it a host of pragmatic factors that inform content ascription, according to the pragmatist picture. The fact that, as the content pragmatist claims, the explanatory gloss in representational terms will never be eliminated from cognitive science (though it may undergo changes as the science progresses), does not entail that content corresponds to some property that cognitive systems, and their internal states objectively possess.

Thus the collapse into content objectivism and robust reductionism is avoided. At the same time, collapse into eliminativism is also averted: the *explananda* of cognitive science require that content be ascribed to (some) of their *explanantia* in order to make intelligible to us how the latter enable the former. We get what the critic feared was not to be had: explanatory ineliminability of content, without ontological commitment — *i.e.* the localised fracture between explanation and ontology that content pragmatism champions.

8.2 The way of the mild reductionist

There are accounts that are deflationary about content in my sense, but that do not lead to content pragmatism. I will present here one of these. It keeps to objectivism about representation and content, but rejects many of the assumptions behind the robust reductionist approach. I call it mild reductionism, or mildIS for short.

Mild reductionism, as I will be using the term, is closely related to the mild realism of Dennett (1981, 1987*a*). Dennett tries to steer a middle course between what he called the industrial-strength Realism about propositional attitudes of Fodor, Dretske, and others, and the eliminativism of Stich and the Churchlands (Dennett 1991) — whilst, *contra* Egan, attempting to keep allegiance to naturalism and objectivism. He claims, briefly, that it is an objective fact that cognitive systems entertain propositional attitudes, such as beliefs and desires. However, and here he departs from industrial-strength Realism, he holds that it is only by taking a specific stance or perspective toward physical systems — the Intentional Stance — that the existence of propositional states becomes discernible (Dennett 1981). Taking the Intentional Stance is justified due to the fact that it enables relatively easy and accurate predictions of the behaviour of some physical systems; ease and accuracy that are obtained by seeing those physical systems as intentional systems — systems that act rationally in their environment in light of what they believe about the world, and what their goals are. Descriptions of such systems — *i.e.* systems to which the Intentional Stance proves to be predictively adequate — in terms of their physical properties, or their design, are possible. But such descriptions are unwieldy and unhelpful when it comes to predicting their behaviour, as they miss higher-level patterns of behaviour that allow capturing useful generalisations in terms of beliefs, desires, and intentional content.

Dennett has been mostly interested in vindicating folk psychology and intentional states through his mild realist view[38]. Given that his target is what I have been calling intentional states and intentional content, Intentional Systems Theory has a marginal bearing on my project. I will not thereby discuss it any further. I will rather apply some of Dennett's insights in developing his mild realist theory of propositional attitudes to the realm of subpersonal states, to the realm, that is, of cognitive states and representational content[39].

Two notions Dennett appeals to and develops in his account are particulary helpful: *abstracta*[40], and real patterns[41]. Let us examine each of these in turn.

The foregoing notion of *abstractum* stems from the work of Reichenbach (1938). I will however use it in an idiosyncratic way, without much regard for how it is cashed

[38] Despite an early adoption of the label 'instrumentalism' to describe his position, with all its pragmatist overtones, Dennett (1987*c*, pp. 71-81) has later on rejected such a label in favour of 'mild realism'. Dennett (1991, p. 51) has expressed some discontentment with the limiting and misleading power of labels.

[39] Dennett (1995, p. 528) claims that the interpretation principles of the Intentional Stance apply also to subpersonal states. However, content-ascription to subpersonal states, by his lights, is parasitic on ascription of folk-psychological states. I am not committed to a view on those lines.

[40] Dennett (1987*c*).

[41] Dennett (1991).

out in Reichenbach's philosophy. As I understand it in the foregoing, *abstracta* capture a complex of objects, states-of-affairs, properties, and processes, and how they are organised — I will call the components of *abstracta*, following Reichenbach (1938, p. 98), their 'internal elements'. Examples of *abstracta* include the political state, and the character of a person[42]. These are (abstract) entities composed of a set of elements of often very different nature, some of which may be *abstracta* in their own right: *e.g.* laws, and their application, government, and their members, citizens, their rights and duties, etc., as well as the complex behaviours and interactions of all of these — and this is admittedly a gross understatement of the complexity of the political state *abstractum*. The *abstracta* that figure as internal elements need themselves be reduced to a set of internal elements that have firm naturalistic credentials, *i.e.* states-of-affairs, events, and processes in the world[43].

Abstracta can be reduced to the elements that compose them, normally by means of a conjunction (or, to use Reichenbach's term, a 'coordination') of those elements, and the relations between them. The meaning of an *abstractum* is equivalent to the meaning of a conjunction of propositions regarding only the internal elements, and their relations. Interestingly, in many cases *abstracta* are composed of a disjunction of coordinated internal elements. This is a consequence of the fact that some *abstracta* are realised by several different sufficient combinations of elements. The *abstractum* 'good weather', for instance, has many different concrete realisations, and therefore can be captured only by a disjunction of coordinated elements. What counts as good weather varies depending on context — *e.g.* season, latitude, closeness to the coast — and even though perhaps it must be sunny in every instance of good weather, how sunny it must be depends on the other coordinated elements, *e.g.* quantity and amount of clouds, wind, temperature, etc.

Abstracta have as much claim to objectivity as the elements that compose them, being complex, often disjunctive arrangements of the latter. The mild reductionist view of representational content that I propose has it that representational contents are *abstracta*. Moreover, they are *abstracta* that bear some resemblance to the *abstractum* 'good weather', in that they are complexes which can only be captured by a disjunction of coordinated internal elements. The claim is that each representational content can be reduced to a (plausibly long, even indefinitely so) disjunction of internal elements, *i.e.* sundry states-of-affairs and processes in the world. I will come back to what types of internal elements compose representational contents below. For now, suffice it to hint at their motley nature. They include, among others, mechanistically-individuated computational structures in the cognitive system, causal relations between the cognitive system, the body and the environment, environmental and bodily context, the history of past causal exchanges with the environment, the evolutionary history of such causal exchanges, and potentially other factors. What the internal elements of each content complex are vary case-by-case, and are strongly context-sensitive.

It is in this sense that 'content' and 'good weather' are alike: there is a large number

[42] Reichenbach (1938, p. 93.)

[43] The nominalistic drive here is clear.

of sundry ways in which a certain content complex can come to exist. *Contra* robust reductionism, there is no privileged exhaustive type of coordination of states-of-affairs and processes such that that type of coordination — *e.g.* in terms of causal-informational relations between cognitive states and the world — and no other, gives rise to representational content. On the contrary, a variety of factors may come into play in different situations and contexts. One factor though must always be there: mechanistically-individuated computational structures internal to the cognitive system. This is the 'narrow', internal factor that must always be present for there to be an instance of representational content. It can never, though, stand alone. And, importantly, the ways in which the computational structure is put to work — *i.e.* which computational states and processes, and combinations thereof play a role in the content complex — may vary considerably when surrounded by different coordinations of other elements, even though these different coordinations give rise to the same instance of the content *abstractum*.

Representational contents are *abstracta*, and *abstracta* are as objective as their internal elements. The internal elements of the content complex, of the content *abstractum*, are objective features of the world — states, processes, and relations such as behaviours and interactions across time, resemblance relations, cognitive, environmental, and historical context. Therefore, representational content is an objective feature of the world.

Robust reductionists are on board with the claim that representational contents are *abstracta*. What sets apart the foregoing from robust reductionism, and what thereby justifies the qualifier in 'mild reductionism', is the way in which content is objective. For robust reductionists, there is a non-disjunctive (or at most very limitedly disjunctive) composition of states-of-affairs and processes that come to compose the content complex. There is, that is to say, a limited and non-indeterminately-disjunctive set of sufficient conditions which, when in place, give rise to representational content.

On the mild reductionist view that I propose, on the other hand, content is determined by a variable, and indeterminately long disjunction of different conjunctions of sundry factors, each giving rise to the same content in different ways. Given the varied role played by the internal factor, computational structure, in this account, there are no identifiable internal states across contexts that bear content essentially. Different combinations of computational states and processes play the part of the internal factor in the content complex depending on the other internal elements that compose the latter. Here, as for the pragmatist view examined above, the primary way to individuate cognitive states is by means of their computational structure. What content structures bear depends on a host of sundry factors, and is thereby too variable and disjunctive to be helpful as the primary means of individuating cognitive states and processes. Individuation by content, though secondary, does play an essential role, again as per the pragmatist: it captures *abstracta* of great complexity, *abstracta* that become scientifically relevant and useful once some physical systems are seen in their active present, ontogenetic, and phylogenetic embeddedness in a changing world.

It will not have slipped the attentive reader's attention that I helped myself to more

than I am entitled to in this last claim. I have put forward so far the view that representational contents are *abstracta* reducible to an indeterminately long disjunction of combinations of factors, or internal elements. However, I have said nothing about what makes those various coordinations of internal elements be part of content complexes. *Abstracta* can be created by sheer stipulation, such as the *bona fide abstractum* 'all bushes in Italy where a rabbit is currently hiding'[44]. But the mild reductionism I advocate would be jeopardised by making representational contents into stipulative or conventional *abstracta*. The content complex would then hinge on the intentions of human beings; it would be an arbitrary way of grouping entities and processes that has no observer-independent foundation. It is crucial for the strand of objectivism that I am exploring that representational contents be non-arbitrary, non-conventional *abstracta*. There must be observer-independent grounds for tying together the variegated and varying factors that give rise to the content complex.

Here the notion of 'real pattern' comes to assistance[45]. 'Pattern' is itself not an easy term to define. One way of understanding patterns is in terms of types of (typically complex) regularities; another closely related one is in terms of relationships between quantities of two or more variables across conditions and over time[46]. Regardless of the best way to characterise patterns, a good criterion to decide if a pattern is at play, following Dennett (1991), is to see whether information about the arrangement of states-of-affairs at hand can be compressed in comparison to a full description of each part of the arrangement. An antonym to 'pattern' is 'randomness' — there is an order to patterns that allows them to be more synthetically described than the disorderliness that marks randomness. An empty chessboard is efficiently described by saying something in the lines of: '8x8 grid of equally-sized squares, alternatively dark- and light-coloured'. A grid of randomly-sized, ramdomly-coloured geometric shapes could never be so efficiently and straightforwardly described — a much longer description would be required in order to capture the nature of that grid, due to its lack of order, due to the absence of a pattern.

To get back to the terminology we have been using, patterns are *abstracta* composed of internal elements tied together by some kind of order or rationale. If they are *real* patterns, the order or rationale that brings together their internal elements is non-conventional, non-stipulative — it is an observer-independent order or rationale found in the world.

It may be countered that patterns are observer-dependent in at least one way: they must be recognised. Dennett (1991, p. 32) admits that the idea of an indiscernible pattern sounds like an oxymoron[47]. However, this is an innocuous type of observer-dependence. What matters is not that patterns be recognised for them to exist, which

[44] I avoid here one of Dennett's (1991) favourite examples — namely the *abstractum* 'the centre of the smallest circle that circumscribes all socks that Daniel Dennett ever lost' — since it is not clear which would be the internal elements of the complex. I do not take *abstracta*, as Dennett (1987*c*, p. 53) does, as "calculation-bound entities or logical constructs".

[45] See Dennett (1991), Nelkin (1994), Haugeland (1998), ter Hark (2001), Burnston (forthcoming).

[46] See Burnston (forthcoming).

[47] In his treatment of real patterns, Haugeland (1998) puts considerable stress on the importance of recognition for the existence of patterns.

would entail strong observer-dependence, but only that they be *candidates* for pattern recognition[48]. Patterns are candidates for pattern recognition precisely because there is an order to them, an order that can be grasped by appropriate observers. Most patterns may never be actually recognised by any observers, due to conceptual and perceptual limitations, or because they are not interesting, or significant to them. Nonetheless, this would make them no less real. As Dennett (1991, p. 34) puts it: "Other creatures with different sense organs, or different interests, might readily perceive patterns that were imperceptible to us ... the patterns would be there all along, but just invisible to us".

The order or rationale that grounds a real pattern may only be grasped, thus making salient the pattern, by means of taking a specific perspective or stance toward the world. By taking that perspective, the hidden order that groups the sundry elements into one *abstractum* is revealed, and the pattern is thus recognised. The only way to recognise some events in the world as instantiating the Caro-Kahn Defence is by seeing them on the backdrop of the rules and strategies of chess. The rules and lore of strategical knowledge of chess is the order or rationale that allows seemingly different physical events in the world — *e.g.* in different games, chessboards, with different pieces, etc. — all to count as members of the same pattern, or *abstractum*, the Caro-Kahn Defence.

My claim is that representational contents are not only *abstracta*: they are real patterns. They are *abstracta* put together in consequence of an order or rationale in the world. In contrast to the rationale behind the Caro-Kahn Defence pattern, in the case of representational contents that order is non-conventional: it is a mind-independent fact about the world. Once the right perspective to grasp that order or rationale is taken, sundry collections of states and processes, current and historical, despite their wildly disjunctive and variable nature, reveal themselves as patterns, as orderly arrangements of elements: as representational contents.

The question then invites itself: what order, and whence?

The phenomena in the world that representational contents capture are the robust, successful interactions between complex organisms and their ever-changing environments, over time and across contexts. Representational contents are the *abstracta* that explain how those robust, successful interactions are made possible by the complex conjunction, in each case, of several different and, at a first glance, unrelated factors of sundry natures — *e.g.* current causal interactions, onto- and phylogenetic history, and so on. What makes those real patterns of behaviour become salient, and reveal themselves as the content pattern is taking the perspective of seeing organisms as largely, though not perfectly, adapted to their historical environments[49]. It is by taking physical systems to be adapted to their environments that their behaviours are revealed as conducive to their maintenance and perpetuation, that their goals become clear, and thereby their success (or lack thereof) in satisfying them[50]. It is the assumption of ad-

[48] Dennett (1991, p. 32.)

[49] I remain neutral on how to identify historical environments. Any plausible way of identifying and type-individuating them will be compatible with the foregoing.

[50] Dennett (1987*b*). The appeal to adaptivity in the foregoing plays an analogous role to the appeal to rationality in Dennett's Intentional Systems Theory.

aptivity that reveals the pattern of robust, goal-directed, successful interaction between organism and world. It is that assumption that makes discernible the objective order, the rationale, that brings together current and historical internal states, causal exchanges, environmental contexts, and yet other factors, to come and compose the content patterns.

In this sense mild reductionism can be seen as a version of interpretational semantics, though it pushes the boundaries of the view considerably. Interpretation here comes from the perspective, or stance, of interpreting the goings-on in the cognitive system under the assumption of adaptivity. Interpretation here is not intentional, as in pragmatism, and neither is it purely mathematical, as in Cummins' interpretational semantics. The interpretation mapping is objective, relying on a rationale that underlies a real pattern in nature, be it recognised or else. Moreover, the interpretation mapping is vastly richer than a mathematical mapping: it takes into account a host of different factors that contribute to internal computational structures being endowed with content in each case. This conception of the interpretation mapping is closer in spirit to Ramsey's than it is to Cummins'. But in contrast to Ramsey's picture, it does not lie at the basis of a robust reductionist account, and it does not cash out interpretation purely in terms of cognitive or behavioural representational use.

In sum, the basic idea behind the mild reductionist flavour of interpretational semantics is to view representational content as capturing certain patterns in nature. These patterns are normally rather complex — they may involve the current state of the cognitive system, its relations with the surroundings, perhaps past interactions that the organism (and its ancestors) had with the environment, and possibly other factors still. Representational contents are *abstracta* that embrace the complicated regularities involving the interaction between organism and world — across contexts, despite disturbing conditions, and so on. It integrates disparate, and apparently independent contributions into a whole that is explanatorily fruitful — it is, as it were, a high-level regularity that makes sense of the coming together of the low-level regularities that compose it.

Representational content, according to mildIS, is a real pattern. It plays an explanatory role in the cognitive sciences, but not as a *mere* explanatory gloss, as for the pragmatist. By shifting much of the burden of explanation to computational structure, the mildly reductionist version of interpretational semantics I defend deflates representational content, but does not make it dependent on our pragmatic purposes. Some internal states, representations, are contentful to the extent that they help compose content *abstracta* by standing in cognitively exploitable relations to things in the world, on the backdrop of other content-relevant factors. That this is so, for the mild reductionist, is an important observer-independent fact about cognitive systems. However, it is, as it were, a 'soft' fact. Representational content does not individuate in a determinate way structures and processes in the cognitive system; it is rather strongly externally context-sensitive — strongly dependent on the environmental embeddedness of the organism at a certain time, as well as on onto- and phylogenetic history — and

strongly internally context-sensitive — strongly dependent on internal cognitive, and bodily context.

The embeddedness of the organism in its environment, its organismic needs, sensory sensitivities and motor capabilities, as well as the state of the environment, determine what the task at hand is, and thus help fix representational content. Those factors are part of the pattern that content is meant to capture. Computational structure mechanistically individuated is another of such factors. What factors come to form the pattern is a matter that may have to be decided case by case. When these (and/or other) factors come together, they give rise to a regularity in nature: the organism behaves in a similar way in structurally similar circumstances. Such a regularity, invisible when we examine the factors responsible for it separately, is brought into sharp focus by having recourse to representational content, in its turn made salient by taking the perspective of adaptivity.

By accepting that the cognitive system deploys such and such representations at a certain point in time, the complex nexus of factors that makes the context-sensitive regularity possible is adequately captured in an efficient and explanatorily powerful manner. It makes clear how (and why) certain computational structures in the cognitive system are deployed in certain situations, and most importantly, it makes clear how come those structures are appropriate to the task — *i.e.* due to the fact that they stand in specific relations with things in the environment, *e.g.* structural resemblance. Representational content reveals the relational nature of cognitive states, relational nature that underpins their capacity to lead to successful behaviour. Such relational nature is captured only when we bring the many factors behind the pattern to the fore.

Now, I have said little about what those factors are, though I have given some suggestions above. I think that we should not try and produce a list of those factors, which would amount to sufficient conditions for representation, and thus lead to a more robustly reductionist theory of representation than what I am willing to embrace[51]. Different factors may be bringing forth the 'representation' pattern in different cases. The same computational structure may have different contents in different contexts, and the same content may be ascribed to different computational structures in different contexts. Nonetheless, that a state has a certain content in a particular situation is a fact about that state, not merely an explanatory gloss. It follows from mildIS that representational content is not the adequate way for the cognitive sciences to individuate internal states — computational structure is. Determinate content-ascription comes on top of that in light of the task at hand, the particular context, and all the others factors forming the content pattern in each case. Computational structures of internal states are definite and context-insensitive, while representational content is fluid, context-sensitive, and variable, as context, broadly understood, is a crucial factor that comes

[51] Nicholas Shea (in progress) is working on such a theory, and hence is keenly interested on delimiting and individuating the factors behind the pattern. Though such an approach is fruitful, I believe that we should be laxer about the relevant factors (and about reductionism), forgoing any attempt to provide a determinate exhaustive list supposed to cover every case. The nature of representational contents as *abstracta* composed of an indeterminately long disjunction of compositions of internal elements, if this mild reductionist view is true, frustrates that attempt.

to form the pattern.

A crucial difference between pragIS and mildIS is that pragmatic considerations play no role in the latter. Among the factors that give rise to the content pattern the interests and aims of cognitive scientists are not to be found. The liberality of second-order resemblance is curbed not by cognitive interpretation, but by context, by the embeddedness of the organism in the environment, and by the other factors, some of which historical, that come to compose the content *abstractum* from case to case.

The mild reductionist about content does not share the deep concern of robust reductionists with indeterminacy problems. In some cases, indeterminate contents will be ascribed due to our epistemic limitations — due to the fact that we lack a full picture of all of the factors composing the pattern, which could decide the matter. In other cases, however, there will be no fact of the matter about which of a series of possibilities is the determinate content of a state — this, I presume, will be more often the case with less complex cognitive systems[52]. In the latter cases, the whole pattern will not justify determinate content ascriptions, though the coming together of the factors forming the content complex will nonetheless be explanatory of behaviour.

For instance, there may be no fact of the matter about whether states in the frog's early visual system represent 'prey there', 'fly there', 'frog food there', or 'dark moving spot there'. But the assumption of adaptivity, and the factors that come to compose the content pattern in this case, are sufficient to explain frogs' behaviour. All of the above content ascriptions are explanatory of their behaviour in their historical environments, and *contra* most of the literature, there is no reason to uphold one over the others. There is no fact of the matter to decide between them, because the content pattern in this case, given the relative simplicity of the cognitive systems of frogs, does not allow (and does not require) such specificity.

Misrepresentation takes place when most, but not all, of the elements of a content pattern are present, and this leads to maladaptive behaviour. For instance, the environment may not be the historical environment for that organism, or some computational error leads to processing mistakes that reflect negatively on behaviour. In this case, the behaviour that will be elicited may fail to make sense under the adaptivity assumption: it may be behaviour that harms, or at least does not contribute, to the survival of the organism. Or else, it may be behaviour that is elicited by a content pattern that, were the factors all there, would have led to adaptive behaviour.

As an example of the first case, a frog may be lured into snapping its tongue when placed in a hostile environment, such as a lab in which humans throw around dark pellets of non-edible material. In this case, there is no answer to the question of what the frog is representing: cases such as these cause a breakdown of the adaptivity assumption. Given onto- and phylogenetic considerations, there is no reason to believe that frogs are adapted to environments other than the historical ones for that species. Of course, in the lab a robust pattern of behaviour reveals itself: frogs snap their tongues at dark pellets of inedible material. However, that pattern of behaviour is not to be explained

[52]See Dennett (1991).

representationally: it is to be explained by a purely computational story, given the impossibility of taking the adaptivity stance in such a case[53].

Second, there is misrepresentation when behaviour is unsuccessful, but the lack of success stems from the instantiation of a pattern which, though underlain by the rationale of adaptivity and thus made salient, is — for some reason or other — not the one that is conducive to adaptivity in a particular context. Naturally-occurring optical illusions are good illustrations of this kind of misrepresentation. In such cases, the adaptivity assumption remains in place, but behaviour is produced that, given sensory and processing limitations, is not adaptive to that situation, though it would have been were the situation different. Recall that the adaptivity assumption has it that organisms are generally adapted to their environments, from which it does not follow that they will always behave adaptively, for instance in conditions that go beyond the bounds of their discriminatory capacities — which are themselves candidate internal elements for the content *abstracta*. Here, the organism may be taken to be representing something else — the content pattern instantiated would be adaptive to a different situation, *e.g.* the one which generates the types of sensory processing carried out by the cognitive system in that instance. These cases are akin to ones in which there are random, blunt errors in computational processing, provided the ensuing computational states and processes are internal elements of some other content pattern. If not, then we are back to the first kind of misrepresentation: there will be no question of what the cognitive system is representing.

Mild reductionism, in brief, is a deflationary view about representation and content that preserves their objectivity, as well as their explanatory role in the cognitive sciences. However, it suggests a substantial shift on what kind of things representations and contents are. Instead of being relatively stable, determinate cognitive structures primarily individuated by their contents, as for robust reductionism and primitivism, representations are context-sensitive combinations of computational vehicles, primarily individuated in terms of the latter. Representational content is not taken as a scientific primitive, but neither is it taken as being reducible to a definite set of sufficient naturalistic factors and relations. Contents are *abstracta* underlain by a rationale in nature; they are real patterns that have as internal elements sundry combinations of factors that may come to compose them in different ways in different circumstances. Mild reductionism thereby sets itself apart from eliminativism, robust reductionism, and primitivism about representation and content. It also features several dissimilarities when compared to its deflationary sibling, content pragmatism. These differences might lead to one view being more promising or satisfactory than the other.

8.3 Pragmatism and mild reductionism compared

As we have seen, both mildIS and pragIS have the tools to address the issues that any theory of content should deal with, such as indeterminacy of content, and misrepres-

[53] See Dennett (1987*a*) for analogous points on cases of irrationality in behaviour, and the consequent breakdown of the rationality assumption.

entation. But it is worthwhile to assess their differences, their theoretical virtues, and shortcomings.

A crucial difference between pragIS and mildIS is that the latter, in contrast to the former, does not make representational content dependent on pragmatic considerations. The content pattern — this complex jumble of contributing factors that leads to significant regularities in the way cognitively complex organisms behave — is out there, in the world, even though it is not a single, bounded object or property somewhere in the brain. Contents are objective features of the world. The content pragmatist, on the other hand, argues that content is assigned to computational structures in light of explanatory aims and interests, being part of an explanatory gloss — objectivism about representation and content is thereby rejected.

The difference between 'explanatory gloss' and 'real pattern' is an interesting one, in which the subtle difference between pragmatism and mild reductionism lies. The content pragmatist recognises that there is a pattern, but denies it is a real pattern in nature, an objective feature of the world. Rather, the pragmatist takes the pattern to arise from our explanatory interests and interpretational capacities — stressing, with Haugeland (1998), the importance of pattern *recognition*. The rationale or order that brings the content pattern to the fore, says pragIS, is constituted by explanatory practices and aims, not some objective, observer-independent order or rationale in the world. There are real patterns relevant for cognition, but those are cashed out in purely computational terms — to which a gloss in representational terms can be added in light of pragmatic considerations.

By preserving the objectivity of content, mild reductionism may be seen as superior to content pragmatism insofar as it is less revisionary, and more in line with mainstream views in philosophy of mind. However, as we have seen throughout this work, such advantage has questionable grounds, and questionable value. It is far from established that content pragmatism is revisionary of the practices of cognitive science[54]. And even if it were, it is not clear how important descriptive accuracy is for a theory of content to be satisfactory. Other theoretical virtues may outweigh a degree of descriptive inaccuracy.

But content pragmatism has two more central shortcomings. First, the grounds for upholding the fracture between explanation and ontology that lies at the basis of the account are shaky, based on the perhaps ill-founded suspicion that representation and content cannot be naturalised. Since this is a possibility, content pragmatism is conceptually coherent, *contra* the critic. But coherence says little about plausibility. Views that do not have to justify such a fracture, such as objectivist views, may have the upper hand if they are independently plausible as theories of content. At any rate, it is comforting to have a back-up plan, a strategy for saving, and justifying the explanatory role of representation in the cognitive sciences even on the off chance that the naturalisation project should fail.

Second, content pragmatism makes representational content parasitic on the inten-

[54] See section 8.1.4.

tional states of human beings, which ground their explanatory interests and practices. But pragIS does not offer an account of intentional content, and it arguably cannot do it using the same tools it uses for explaining representational content. Content pragmatism has an explanatory lacuna at its core: it employs an obscure notion on which it is silent — intentional content — to make sense of representational content. It thereby makes a heterodox, and arguably dubious move: to explain simpler phenomena by means of more complex ones.

These are not decisive defects of the view. But they suggest that theories of content that avoid the explanatory fracture, and the explanatory lacuna that mark content pragmatism may be preferable — provided they are satisfactory in their own right. MildIS may be one such theory. It is objectivist throughout, and therefore needs to effect no fracture between explanation and ontology. Its account of representational content, moreover, does not appeal to more obscure or complex notions, such as intentional content. I believe that it should be preferred to pragIS.

It may be argued that mild reductionism is too unconstrained and epistemically opaque. Content, being a complex and variegated affair, becomes difficult to ascribe with certainty, given all the factors that have to be taken into account in each case, which may moreover not be generalisable to other cases and contexts. This might be taken as jeopardising the explanatory usefulness of the notion in cognitive science.

I do not think that the epistemic opacity that follows from mildIS is damaging. It is to be expected that a notion so complex as representational content should lead to difficulties pinpointing with confidence what contents are at play in each case. Choices in terms of explanatory power and aims may be made in order to help content ascription, but these factors do not constitute the content pattern itself. And there may be cases in which there is no determinate content to start with, so that attempts to ascribe definite contents would only approximate the real pattern.

Compare with the political state *abstractum*: it is often difficult to determine whether some states are democracies. But we can reach good enough judgements on the basis of non-exhaustive knowledge of the internal elements that compose the *abstractum*. Something similar is true of representational content. Though we may expect considerable epistemic opacity, the notion is nonetheless useful and explanatory. For we can have access to several elements of the pattern, and therefore make informed hypotheses about its nature. And investigation may reveal further elements, changing or refining our understanding of what contents are at play in specific cases.

A further worry regards the capacity of mildIS to allow useful generalisations. If content is strongly context-sensitive, and constituted by an indefinite disjunction of sets of sufficient factors, it might seem that any explanation in terms of content will not generalise beyond single instances of behaviour, hurting the explanatory value of resorting to representation. I do not think it follows from mildIS that generalisations are impossible. For most explanatory purposes, relatively coarse-grained content ascriptions suffice, so that the subtle differences in content that might derive from slightly different contexts can be ignored. Moreover, given that content patterns involve diachronic

considerations, different instances of behaviour will fall under the same pattern, allowing explanations that extrapolate from current instances, to past, and perhaps future ones. The explanatory purchase of representational content is preserved, despite the complexity, and consequent epistemic opacity that ensues.

In sum, I believe that there are reasons to prefer mild reductionism to pragmatism, though these reasons are not decisive, based on some theoretical virtues that the former has, when compared to the latter.

8.4 *Coda* — Workings and roles: neural reuse and deflated representation

The two deflationary paths that I have put forward above, and especially the mild reductionist version, entail some empirical claims about the nature of the internal representational vehicles — that is to say, the internal computational states to which representational contents are ascribed, be it for pragmatic reasons, in pragIS, or by means of being part of the relevant real pattern, in mildIS. These views entail that representational vehicles, and perhaps the realisers of cognitive states and processes in general, are in complex, dynamic, and context-sensitive mapping relations with their functional and representational characterisations. The deflationary account I propose has it that the relationship between contents and computational vehicles is many-to-many: there may be cases in which one computational vehicle has or is ascribed several different contents, depending on context; and there may be cases in which several computational vehicles, in similar or different contexts, have or are ascribed the same content. These apparently heterodox claims might appear too big a bullet to bite, unless some further motivation is provided to back them up.

In this section, I will use recent work in the field of cognitive ontology to argue that the empirical claims that follow from my deflationary view are not only conceptually possible, but also scientifically plausible. My view has good scientific credentials insofar as it dovetails nicely with current discussions on how best to understand and describe the functional structure of the brain, and how it relates to its physical structure. These discussions aim at providing an adequate 'cognitive ontology' to the cognitive sciences.

Ontologies[55] are 'systematic descriptions' of "structure-function relations whereby structures predict functions and functions predict structures"[56]. Cognitive ontologies are in the business of figuring out how a) to carve cognitive systems both functionally and structurally, and b) to map functionally-individuated states and processes at the cognitive level (*e.g.* face recognition, verbal articulation) into structurally-individuated states and processes at the physiological level (*e.g.* brain areas, networks, neural circuits)[57]. Such mapping relations, as many in the debate have pointed out, are likely not one-to-one, *contra* the still in many respects dominant localist paradigm[58]. There

[55] Note that the term 'ontology' is not used here in its traditional philosophical sense.

[56] Price & Friston (2005, p. 263.)

[57] Anderson (2015) provides a brief introduction to the debate about cognitive ontologies.

[58] See, for illustration and discussion, Noppeney et al. (2004), McIntosh (2004), Price & Friston

are reasons to believe that complex cognitive systems such as the human one feature brain areas that are pluripotential — one area subserves many cognitive functions — and degenerate — many areas are sufficient for the same cognitive function — which make implausible the one-to-one mappings the localist paradigm would have wanted[59]. In contrast, we should expect and look for many-to-many mappings between cognitive functions and brain structures.

The issues and questions that exercise theorists in cognitive ontology are related, but not equivalent, to the issues and questions that arise from a treatment of cognitive representation. But the former bear on the latter. Cognitive ontologies concern cognitive functions posited in light of functional decompositions of cognitive tasks. These cognitive subprocesses, *e.g.* facial recognition, are at a level of grain considerably coarser than the one that would individuate specific representational vehicles, and their contents. Nevertheless, pluripotentiality and degeneracy at this level of description of cognitive systems bear on questions relating to representational vehicles. For if one and the same area, in the same cognitive system, and over a timeframe that bars significant structural change, is able to participate in cognitive functions of very different kinds, as the empirical evidence suggests, there is reason to believe that the representational vehicles that it realises or contributes to realising vary in their contents, if not even in their functional individuation. Analogously, if one and the same cognitive function can be implemented in structurally different areas, in which the structure of local neural circuits differ, then there is reason to believe that the representations relevant for that function will be realised in those structurally different areas, perhaps in different ways[60].

The debate on how to go about in looking for the right cognitive ontologies for human cognition is quite recent, and there is as yet much work to be done, both conceptual and empirical. Different positions can be delineated in the contemporary debate[61]. A conservative attitude has it that the psychological categories that we apply in our study of cognition will be revised in light of empirical work in neuropsychology, but most of our psychological constructs will be preserved, and more general functional characterisations will lead to one-to-one function-structure mappings. A moderate take accepts that work in neuropsychology may lead to a more marked revision of our psychological constructs than that predicted by the conservative: our current constructs may be in some cases merged with each other, in other cases split up, and in others still eliminated — nonetheless, one-to-one structure-function mappings will typically be possible. More radical positions claim that we will eventually end up with constructs that are sub-

(2005), Anderson (2010), Figdor (2010), Klein (2012), Rathkopf (2013), McCraffrey (2015), Bergeron (2016), Burnston (2016*b*).

[59] For reviews of the empirical evidence, see Price & Friston (2005), Anderson (2010), Figdor (2010).

[60] As Noppeney et al. (2004), Price & Friston (2005), Figdor (2010) point out, function-structure mappings, as well as issues relating to pluripotentiality and degeneracy, vary according to level of description. Two different areas may be sufficient for the task 'reading out loud', and thus be considered degenerate. However, they may be implementing different strategies, *i.e.* spelling-to-sound *vs.* lexico-semantic memory, that lead to the same behavioural outcome. So at a lower level of description, those areas are not degenerate relative to one another, since each is not able to perform the specific function the other performs. For our purposes, degeneracy at a rather low level of description is the most relevant case.

[61] Anderson (2015, pp. 70ff.).

stantially different from the ones we have now, and that one-to-one structure-function mappings, far from the normal case, will be atypical. They hold that neuroscientific evidence motivates (and will motivate) a fundamental revision of the ontology of cognition[62] — one in which functional characterisations of brain regions will involve functions that have little to do with the psychological constructs we posit in neuropsychological research nowadays.

One of the most well-argued positions in the radical end of the spectrum has been put forward by Michael L. Anderson in a series of papers (2007*a*, 2007*b*, 2010, 2016), and in a recent book (2014). According to this view, the Massive Redeployment Hypothesis, the brain employs neural circuits in a variety of different cognitive tasks, such that areas of the brain come to play sundry cognitive roles. Neural reuse theories, of which the Massive Redeployment Thesis is a version, have it that "low-level neural circuits are used and reused for various purposes in different cognitive and task domains"[63]. The basic idea is that cognitive functions are realised by different combinations of many neural circuits, each of which has fixed properties. Anderson (2010, 2014, 2016) sees his version of the neural reuse hypothesis as incorporating, and expanding on the project of 4E cognition — *i.e.* cognition as embedded, embodied, enacted, and extended. Though this application of the view is interesting, there are more moderate and less revisionary possibilities as to how to develop the idea that neural reuse is a widespread property of the brain[64]. On a relatively moderate understanding of neural reuse, a more precise revision of our cognitive ontology will focus on the computational profile of neural circuits at the lowest, context-insensitive level of description of the cognitive system. Neural circuits have definite computational capacities, and different cognitive functions can be realised by different combinations of neural circuits in light of those capacities.

A crucial distinction is in order to make sense of the proposal: that between the 'working' of a neural circuit, and its 'use' (or 'role') in a specific neural and cognitive context[65]. As Anderson (2010, p. 252) nicely summarises, neural reuse models

> make a strong distinction between a "working" — whatever specific computational contribution local anatomical circuits make to overall function — and a "use", the cognitive purpose to which the working is put in any individual case. For neural reuse theories, anatomical sites have a fixed working, but many different uses.

The workings of a neural circuit are the "low-level computational operations" it performs, and which can be exploited in the bringing about of different "higher-level cognitive uses"[66]. Workings are context-insensitive; they are the intrinsic computations that a neural circuit performs. The cognitive uses to which they are put depend on

[62] For criticism of the putative evidence so far available to substantiate radical views, see Kaplan & Craver (2016), Shine et al. (2016).

[63] Anderson (2010, p. 246.)

[64] See Bergeron (2007, 2016), Shine et al. (2016).

[65] Bergeron (2007, 2016), Anderson (2010). Anderson (2016) downsizes the importance of the distinction, embracing a more radical view according to which there will be cases in which workings cannot be determined independently of uses. See also Burnston (2016*a*).

[66] Bergeron (2016, p. 819.)

the neural context in which they are inserted from case to case. Their computational contribution to each cognitive use remains invariant, but due to the different combinations of computations recruited by different cognitive functions, the cognitive roles those neural circuits play, and thereby the overall computations performed, depend on which other connected neural circuits are at work in each case — *i.e.* which subset of the computational structures of the cognitive system is activated in a specific context — as well as on the source of the inputs received and processed[67]. A similar position on neural reuse is championed by Carandini & Heeger (2012, p. 51), who claim that

> physiological and behavioural evidence suggests that canonical neural computations exist — standard computational modules that apply the same fundamental operations in a variety of contexts. A canonical neural computation can rely on diverse circuits and mechanisms, and different brain regions or different species may implement it with different available components.

Candidate fundamental, or 'canonical' neural computations, include: exponentiation, linear filtering, and normalisation[68]. But there may be many others. These are computations implemented in different neural structures and areas of the brain, which play a role in bringing about different cognitive functions. There is evidence that neural circuits computing normalisation play a role in cognitive functions such as olfactory, auditory, and visual sensory processing — making contributions at different stages of processing in the visual system, from the retina, to V1 and MT, supporting different cognitive functions in each case (*e.g.* object recognition) — and attention modulation[69].

The relevance of these considerations to the deflationary view of representation should be clear. If a version of the neural reuse hypothesis on these lines is correct, then the basic and primary way of individuating internal states relevant for cognition is in terms of their computational structure, formed by various combinations of computationally-specialised neural circuits. Representation comes on top of that computational description, when the aim is to explain higher-level cognitive uses — the cognitive functions to which those networks of circuits are contributing. Given that, according to neural reuse theories, the same neural circuits, with their fixed workings, are at play in many different types of cognitive functions, the representational roles they

[67] Burnston (2016*a*) has put into doubt the thesis that workings are context-insensitive, thesis that he dubs 'computational absolutism', on the grounds that current computational models ascribe different computational profiles to brain area MT depending on which cognitive function is being performed. He argues that workings vary according to context, as the case study of area MT would suggest. Though this strong contextualist view is not incompatible with the position I am defending here — since mechanistic explanation, and thereby mechanistically-individuated computational structure hinge on what the *explanandum* cognitive function is — Burnston's arguments do not justify the rejection of computational absolutism. His reliance on current computational models weakens his claims, for there is no reason to believe that our current and rather coarse computational models, even though predictive, actually mirror the computations being performed by specific neural circuits. Moreover, often those computational models are meant to capture different levels of description of the system, using different computational paradigms. Diversity in models is to be expected, given their different aims. Moreover, even sticking to the basic level of workings, it may be that different cognitive tasks recruit different neural circuits, and combinations thereof, in MT, thus giving rise to different computational profiles.

[68] See also Shine et al. (2016).

[69] See Carandini & Heeger (2012) for a review of the evidence.

play, as well as their computational junctures themselves, vary in a context-sensitive way in light of neural, bodily, and environmental context — as my deflationary view would have it.

If a relatively moderate neural reuse theory is true, there is little motivation to see representation as the robust reductionist would want it: as definable, repeatable cognitive structures that essentially bear their contents, which they get by standing in specifiable naturalistic sets of relations to entities in the body and the world. That picture would rather suggest that representation may be a much more flexible, context-sensitive, and variable affair — something more in the lines of the deflationary, computation-based theory I am offering, regardless of whether one opts for its pragmatic or mild reductionist flavour.

Whether this sort of hypothesis will pan out is a largely empirical matter. My aim was just to put at least some empirical flesh around my philosophical bones. That a philosophical thesis should make empirical predictions, and its interest depend at least in part on their coming out true, is, I take, as it should be — at least when it comes to the philosophy of cognitive science.

Concluding remarks

We have reached the core of my project. I examined extant versions of interpretational semantics, their advantages and shortcomings, and argued that they should be supplemented with the robust notion of concrete computation offered by the mechanistic view defended in Part II. That move allowed me to put forward a theory of representation that draws from both interpretational semantics and structural representation. Second-order resemblance plays an important role in at least one type of representation, while the notion of content is deflated, and with it many philosophical worries.

As a final, and somewhat tangential, remark, the theory I am here advocating, in its two guises, is nicely wedded to pluralism about representation. Second-order resemblance might be one of the ways in which computational structures lead to successful behaviour. The possibility is open that there may be other relations between internal states and the world that are equally explanatorily useful. My focus on theories of representation based on second-order resemblance is motivated by the importance and fruitfulness of the notion of structural representation, of the representation-as-model model[70], in the cognitive sciences. There might be other models of representation that are explanatorily fruitful, and compatible with the foregoing account.

In the picture I offer, no unified account should be sought. Representational content captures a variety of different, fluid cognitive states and processes that mediate complex behaviour. How this mediation is effected may vary wildly in different situations and tasks. One of them, I believe, is based on second-order resemblance relations between computational structures, and target domains in the world. Ascription of representational content to such internal states and processes is a way of generalising over the diversity of physical and computational processes in the cognitive system that lead to behavioural success in specific situations. This deflated understanding of representation becomes possible once we have a robust notion of concrete computation. The account, in its two flavours, is not merely instrumentalist, given the reliance on computational structures and computational mechanisms. For the same reason, neither is it eliminativist. Representational content, even in the pragmatist flavour, keeps its explanatory importance, and cannot be eliminated from scientific discourse. The place of representation in the cognitive sciences is preserved, unencumbered by metaphysical burdens.

[70] Godfrey-Smith (2009*a*).

Bibliography

Adams, F. (2010), 'Why we still need a mark of the cognitive', *Cognitive Systems Research* **11**(4), 324–331.

Andersen, H. (2014*a*), 'A field guide to mechanisms: Part I', *Philosophy Compass* **9**(4), 274–283.

Andersen, H. (2014*b*), 'A field guide to mechanisms: Part II', *Philosophy Compass* **9**(4), 284–293.

Anderson, M. L. (2007*a*), 'Massive redeployment, exaptation, and the functional integration of cognitive operations', *Synthese* **159**, 329–345.

Anderson, M. L. (2007*b*), 'The massive redeployment hypothesis and the functional topography of the brain', *Philosophical Psychology* **20**(2), 143–174.

Anderson, M. L. (2010), 'Neural reuse: a fundamental organizational principle of the brain', *Behavioral and Brain Sciences* **33**(245-313).

Anderson, M. L. (2014), *After Phrenology: Neural Reuse and the Interactive Brain*, MIT Press.

Anderson, M. L. (2015), 'Mining the brain for a new taxonomy of the mind', *Philosophy Compass* **10**(1), 68–77.

Anderson, M. L. (2016), 'Précis of after phrenology: Neural reuse and the interactive brain', *Behavioral and Brain Sciences* .

Artiga, M. (2014), 'Prinz's naturalistic theory of intentional content', *Crítica, Revista Hispanoamericana de Filosofía* **46**(136), 69–86.

Artiga, M. & Martínez, M. (2016), 'The organizational account of function is an etiological account of function', *Acta Biotheoretica* **64**(2), 105–117.

Bartels, A. (2006), 'Defending the structural concept of representation', *Theoria* **21**(55), 7–19.

Bechtel, W. (2016), 'Investigating neural representations: the tale of place cells', *Synthese* **193**, 1287–1321.

Bergeron, V. (2007), 'Anatomical and functional modularity in cognitive science: Shifting the focus', *Philosophical psychology* **20**(2), 175–195.

Bergeron, V. (2016), 'Functional independence and cognitive architecture', *British Journal for the Philosophy of Science* **67**, 817–836.

Bigelow, J. & Pargetter, R. (1987), 'Functions', *The Journal of Philosophy* **84**(4), 181–196.

Blackburn, S. (2010), 'The presidential address: The steps from doing to saying', *Proceedings of the Aristotelian Society* **110**, 1–13.

Block, N. (1986), 'Advertisement for a semantics for psychology', *Midwest Studies in Philosophy* **X**(1), 615–78.

Boorse, C. (1976), 'Wright on functions', *The Philosophical Review* **85**(1), 70–86.

Brown, J. W. (2014), 'The tale of the neuroscients and the computer: why mechanistic theory matters', *Frontiers in Neuroscience* **8**.

Buller, D. J. (1998), 'Etiological theories of function: a geographical survey', *Biology and Philosophy* **13**, 505–527.

Burge, T. (2010), *Origins of Objectivity*, Oxford University Press.

Burnston, D. C. (2016*a*), 'Computational neuroscience and localized neural function', *Synthese* **193**, 3741–3762.

Burnston, D. C. (2016*b*), 'A contextual approach to functional localisation in the brain', *Biology and Philosophy* **31**(4), 527–550.

Burnston, D. C. (forthcoming), 'Real patterns in biological explanation', *Proceedings of the Philosophy*

of Science Association .

Carandini, M. & Heeger, D. J. (2012), 'Normalization as a canonical neural computation', *Nature Neuroscience* **13**, 51–62.

Cash, M. (2009), 'Normativity is the mother of intention: Wittgenstein, normative practices and neurological representations', *New Ideas in Psychology* **27**, 133–47.

Chalmers, D. J. (1995), 'On implementing a computation', *Minds and Machines* **4**, 391–402.

Chalmers, D. J. (1996), 'Does a rock implement every finite-state automaton?', *Synthese* **108**, 309–333.

Chalmers, D. J. (2011), 'A computational foundation for the study of cognition', *Journal of Cognitive Science* **12**(4), 323–357.

Chalmers, D. J. (2012), 'The varieties of computation: a reply', *Journal of Cognitive Science* **13**, 211–248.

Chomsky, N. (1995), 'Language and nature', *Mind* **104**(413), 1–61.

Chrisley, R. L. (1995), 'Why everything doesn't realise every computation', *Minds and Machines* **4**, 403–420.

Churchland, P. M. (1981), 'Eliminative materialism and the propositional attitudes', *Journal of Philosophy* **78**, 67–90.

Churchland, P. M. (1998), 'Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered', *The Journal of Philosophy* **95**(1), 5–32.

Churchland, P. M. (2012), *Plato's camera : how the physical brain captures a landscape of abstract universals*, MIT Press.

Churchland, P. S., Koch, C. & Sejnowski, T. J. (1990), What is computational neuroscience?, *in* E. L. Schwartz, ed., 'Computational Neuroscience', The MIT Press.

Clark, A. & Toribio, J. (1994), 'Doing without representing?', *Synthese* **101**, 401–431.

Coelho Mollo, D. (2015), 'Being clear on content', *Philosophia* **43**(3), 687–699.

Copeland, B. J. (1996), 'What is computation?', *Synthese* **108**, 335–359.

Craver, C. F. (2001), 'Role functions, mechanisms, and hierarchy', *Philosophy of Science* **68**, 53–74.

Craver, C. F. (2006), 'When mechanistic models explain', *Synthese* **153**, 355–376.

Craver, C. F. (2007), *Explaining the Brain*, Oxford University Press.

Craver, C. F. (2013), Functions and mechanisms: a perspectivalist view, *in* P. Huneman, ed., 'Functions: Selection and Mechanisms', Springer.

Craver, C. F. (2014), The ontic account of scientific explanation, *in* M. I. Kaiser, O. R. Scholz, D. Plenge & A. Hüttemann, eds, 'Explanation in the Special Sciences: The case of Biology and History', Springer Netherlands.

Craver, C. F. & Tabery, J. (2016), Mechanisms in science, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy (Spring 2016 Edition)'.
URL: *http://plato.stanford.edu/archives/spr2016/entries/science-mechanisms/*

Cummins, R., Blackmon, J., Byrd, D., Lee, A. & Roth, M. (2010/2006), Representation and unexploited content, *in* 'The World in the Head', Oxford University Press, pp. 120–134.

Cummins, R. C. (1975), 'Functional analysis', *Journal of Philosophy* **72**(20), 741–765.

Cummins, R. C. (1983), *The Nature of Psychological Explanation*, MIT Press.

Cummins, R. C. (1989), *Meaning and Mental Representation*, MIT Press.

Cummins, R. C. (1996), *Representations, Targets, and Attitudes*, MIT Press.

Cummins, R. C. & Schwarz, G. (1991), Connectionism, computation, and cognition, *in* T. E. Horgan & J. L. Tienson, eds, 'Connectionism and the Philosophy of Mind', Kluwer Academic Publishers.

Dennett, D. C. (1971), 'Intentional systems', *The Journal of Philosophy* **68**(4), 87–106.

Dennett, D. C. (1981), True believers: The intentional strategy and why it works, *in* A. F. Heath, ed., 'Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford', Clarendon Press, pp. 150–167.

Dennett, D. C. (1987*a*), *The Intentional Stance*, The MIT Press.

Dennett, D. C. (1987*b*), Intentional systems in cognitive ethology: the "panglossian paradigm" defended, *in* 'The Intentional Stance', MIT Press, pp. 237–286.

Dennett, D. C. (1987*c*), Three kinds of intentional psychology, *in* 'The Intentional Stance', MIT Press,

pp. 43–81.

Dennett, D. C. (1991), 'Real patterns', *The Journal of Philosophy* **88**(1), 27–51.

Dennett, D. C. (1995), 'Get real', *Philosophical Topics* **22**(1-2), 505–568.

Dewhurst, J. (2016), 'Individuation without representation', *British Journal for the Philosophy of Science* .

Dresner, E. (2010), 'Measurement-theoretic representation and computation-theoretic realization', *The Journal of Philosophy* **CVII**(6), 275–292.

Dretske, F. (1981), *Knowledge and the Flow of Information*, Basil Blackwell.

Dretske, F. (1986), Misrepresentation, *in* R. Bogdan, ed., 'Belief: Form, Content, and Function', Oxford University Press, pp. 17–36.

Dretske, F. (1988), *Explaining Behavior*, MIT Press.

Dunlop, C. E. M. (2004), 'Mentalese semantics and the naturalized mind', *Philosophical Psychology* **17**(1), 77–94.

Egan, F. (1995), 'Computation and content', *Philosophical Review* **104**(2), 181–203.

Egan, F. (1999), 'In defence of narrow mindedness', *Mind & Language* **14**(2), 177–194.

Egan, F. (2009), Is there a role for representational content in scientific psychology?, *in* D. Murphy & M. A. Bishop, eds, 'Stich and His Critics', Wiley-Blackwell.

Egan, F. (2010), 'Computational models: a modest role for content', *Studies in History and Philosophy of Science Part A* **41**(3), 253–259.

Egan, F. (2012), 'Metaphysics and computational cognitive science: Let's not let the tail wag the dog', *Journal of Cognitive Science* **13**, 39–49.

Egan, F. (2014*a*), 'Explaining representation: a reply to Matthen', *Philosophical Studies* **170**, 137–142.

Egan, F. (2014*b*), 'How to think about mental content', *Philosophical Studies* **170**, 115–135.

Egan, F. (2015), 'A deflationary account of mental representation'. Presentation at the conference 'Mental Representations, the foundation of Cognitive Science?', Ruhr-Universitaet Bochum, 21-23 September 2015. Available online at: http://www.youtube.com/watch?v=Sum5HNMibnU.

Egan, F. (forthcoming), Function-theoretic explanation and the search for neural mechanisms, *in* D. M. Kaplan, ed., 'Integrating Mind and Brain Science: Mechanistic Perspectives and Beyond', Oxford University Press.

Eliasmith, C. (2000), How Neurons Mean: a neurocomputational theory of representational content, PhD thesis, Washington University in St. Louis.

Eliasmith, C. (2005), 'A new perspective on representational problems', *Journal of Cognitive Science* **6**, 97–123.

Figdor, C. (2010), 'Neuroscience and the multiple realization of cognitive functions', *Philosophy of Science* **77**, 419–456.

Fodor, J. A. (1975), *The Language of Thought*, Harvard University Press.

Fodor, J. A. (1981), *Representations: Philosophical Essays on the Foundations of Cognitive Science*, The Harvester Press.

Fodor, J. A. (1984), 'Semantics, Wisconsin Style', *Synthese* **59**(3), 231–250.

Fodor, J. A. (1987), *Psychosemantics*, MIT Press.

Fodor, J. A. (1990), *A Theory of Content and Other Essays*, MIT Press.

Fodor, J. A. (2000), *The Mind Doesn't Work that Way*, MIT Press.

Fodor, J. A. (2008), *LOT 2: The Language of Thought Revisited*, Oxford University Press.

Fodor, J. A. & Lepore, E. (1992), *Holism: a shopper's guide*, Blackwell Publishing.

Fodor, J. A. & Lepore, E. (1999), 'All at sea in semantic space: Churchland on meaning similarity', *The Journal of Philosophy* **96**(8), 381–403.

Fresco, N. (2014), *Physical Computation and Cognitive Science*, Springer.

Fresco, N. (2015), 'Objective computation versus subjective computation', *Erkenntnis* **80**(5), 1031–1053.

Fresco, N. & Primiero, G. (2013), 'Miscomputation', *Philosophy and Technology* **26**(3), 253–272.

Fresco, N., Wolf, M. J. & Copeland, J. B. (2016), On the indeterminacy of computation, *in* 'Methodological Issues in Philosophy of Computer Science Symposium. The 2016 Annual Meeting of the

International Association for Computing and Philosophy, University of Ferrara, Italy'.

Fresco, N., Wolf, M. J. & Copeland, J. B. (forthcoming), 'The indeterminacy of computation: computational explanations and neural mechanisms'.

Gallistel, C. R. (1990), *The Organization of Learning*, The MIT Press.

Garson, J. (2011), 'Selected effects and causal role functions in the brain: the case for an etiological approach to neuroscience', *Biology and Philosophy* **26**(4), 547–565.

Garson, J. (2013), 'The functional sense of mechanism', *Philosophy of Science* **80**, 317–333.

Garson, J. & Piccinini, G. (2014), 'Functions must be performed at appropriate rates in appropriate situations', *British Journal for the Philosophy of Science* **65**, 1–20.

Gates, G. (1996), 'The price of information', *Synthese* **107**(3), 325–347.

Glennan, S. S. (1996), 'Mechanisms and the nature of causation', *Erkenntnis* **44**, 49–71.

Glennan, S. S. (2010*a*), 'Ephemeral mechanisms and historical explanation', *Erkenntnis* **72**, 251–266.

Glennan, S. S. (2010*b*), Mechanisms, *in* H. Beebee, C. Hitchcock & P. Menzies, eds, 'The Oxford Handbook of Causation', Oxford University Press.

Godfrey-Smith, P. (1993), 'Functions: Consensus withouth unity', *Pacific Philosophical Quarterly* **74**(3), 196–208.

Godfrey-Smith, P. (1994), 'A modern history theory of functions', *Noûs* **28**(3), 344–362.

Godfrey-Smith, P. (2006), Mental representation, naturalism, and teleosemantics, *in* D. Papineau & G. Macdonald, eds, 'Teleosemantics: New Philosophical Essays', Clarendon Press.

Godfrey-Smith, P. (2009*a*), Representationalism reconsidered, *in* D. Murphy & M. A. Bishop, eds, 'Stich and His Critics', Wiley-Blackwell.

Godfrey-Smith, P. (2009*b*), 'Triviality arguments against functionalism', *Philosophical Studies* **145**, 273–295.

Goodman, N. (1976), *Languages of Art: An Approach to a Theory of Symbols*, Hackett Publishing.

Grush, R. (2001), The semantic challenge to computational neuroscience, *in* P. M. Peter K. Machamer & R. Grush, eds, 'Theory and Method in the Neurosciences', University of Pittsburgh Press.

Grush, R. (2004), 'The emulation theory of representation: Motor control, imagery, and perception', *Behavioural and Brain Sciences* **27**, 377–442.

Haimovici, S. (2013), 'A problem for the mechanistic account of computation', *Journal of Cognitive Science* **14**, 151–181.

Halina, M. (forthcoming), Mechanistic explanation and its limits, *in* S. Glennan & P. Illari, eds, 'Routledge Handbook of Philosophy of Mechanisms', Routledge.

Harman, G. (1982), 'Conceptual role semantics', *Notre Dame Journal of Formal Logic* **23**(2), 242–56.

Haugeland, J. (1981), Semantic engines: An introduction to mind design, *in* 'Mind Design', MIT Press.

Haugeland, J. (1985), *Artificial Intelligence: the very idea*, MIT Press.

Haugeland, J. (1990), 'The Intentionality All-Stars', *Philosophical Perspectives* **4**, 383–427.

Haugeland, J. (1998), Pattern and being, *in* 'Having Thought: Essays in the Metaphysics of Mind', Harvard University Press, pp. 267–290.

Hutto, D. D. & Myin, E. (2013), *Radicalizing Enactivism: Basic minds without content*, The MIT Press.

Illari, P. M. & Williamson, J. (2012), 'What is a mechanism? Thinking about mechanisms across the sciences', *European Journal for the Philosophy of Science* **2**, 119–135.

Isaac, A. M. C. (2012), 'Objective similarity and mental representation', *Australasian Journal of Philosophy* **91**(4), 683–704.

Johnson-Laird, P. N. (1983), *Mental Models*, Cambrigde University Press.

Kaplan, D. M. & Craver, C. F. (2016), 'A registration problem for functional fingerprinting', *Behavioral and Brain Sciences* .

Klein, C. (2008), 'Dispositional implementation solves the superfluous structure problem', *Synthese* **165**, 141–153.

Klein, C. (2012), 'Cognitive ontology and region- versus network-oriented analyses', *Philosophy of Science* **79**(5), 952–960.

Krohs, U. (2009), 'Functions as based on a concept of general design', *Synthese* **166**(1), 69–89.

Ladyman, J. (2009), 'What does it mean to say that a physical system implements a computation?', *Theoretical Computer Science* **410**, 376–383.

Levy, A. (2013), 'Three kinds of new mechanism', *Biology and Philosophy* **28**, 99–114.

Levy, A. & Bechtel, W. (2013), 'Abstraction and the organization of mechanisms', *Philosophy of Science* **80**, 241–261.

Machamer, P., Darden, L. & Craver, C. (2000), 'Thinking about mechanisms', *Philosophy of Science* **67**, 1–25.

Markman, A. B. & Dietrich, E. (2000), 'Extending the classical view of representation', *Trends in Cognitive Science* **4**(12), 470–475.

Marr, D. (1982), *Vision*, Freeman.

Matthews, R. J. (2011), 'Measurement- theoretic accounts of propositional attitudes', *Philosophy Compass* **6**(11), 828–841.

Matthews, R. J. & Dresner, E. (2016), 'Measurement and computational skepticism', *Noûs* .

Maudlin, T. (1989), 'Computation and consciousness', *The Journal of Philosophy* **86**(8), 407–432.

McCraffrey, J. B. (2015), 'The brain's heterogeneous functional landscape', *Philosophy of Science* **82**, 1010–1022.

McIntosh, A. R. (2004), 'Contexts and catalysts: a resolution of the localization and integration of function in the brain', *Neuroinformatics* **2**, 175–181.

McLendon, H. J. (1955), 'Uses of similarity of structure in contemporary philosophy', *Mind* **64**(253), 79–95.

Milkowski, M. (2011), 'Beyond formal structure: A mechanistic perspective on computation and implementation', *Journal of Cognitive Science* **12**, 359–379.

Milkowski, M. (2012), 'Is computation based on interpretation?', *Semiotica* **188**(1/4), 219–228.

Milkowski, M. (2013), *Explaining the Computational Mind*, The MIT Press.

Milkowski, M. (2016), 'Function and causal relevance of content', *New Ideas in Psychology* **40**, 94–102.

Millikan, R. G. (1984), *Language, Thought, and Other Biological Categories: New Foundations for Realism*, MIT Press.

Millikan, R. G. (1989*a*), 'Biosemantics', *The Journal of Philosophy* **86**(6), 281–297.

Millikan, R. G. (1989*b*), 'In defense of proper functions', *Philosophy of Science* **56**(2), 288–302.

Millikan, R. G. (2000), Naturalizing intentionality, *in* B. Elevitch, ed., 'Proceedings of the Twentieth World Congress of Philosophy', pp. 83–90.

Millikan, R. G. (2004), *Varieties of Meaning: the 2002 Jean Nicod Lectures*, MIT Press.

Morgan, A. (2014), 'Representations gone mental', *Synthese* **191**, 213–244.

Moser, E. I., Kropff, E. & Moser, M.-B. (2008), 'Place cells, grid cells, and the brain's spatial representation system', *Annual Review of Neuroscience* **31**, 69–89.

Moss, L. (2012), 'Is the philosophy of mechanism philosophy enough?', *Studies in History and Philosophy of Biological and Biomedical Sciences* **43**, 164–172.

Mossio, M., Saborido, C. & Moreno, A. (2009), 'An organizational account of biological functions', *British Journal for the Philosophy of Science* **60**, 813–841.

Neander, K. (1991), 'Functions as selected effects: The conceptual analyst's defence', *Philosophy of Science* **58**(2), 168–184.

Neander, K. (2012), Teleological theories of mental content, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy (Spring 2012 Edition)', URL = <https://plato.stanford.edu/archives/spr2012/entries/content-teleological/>.

Neander, K. (2013), Toward an informational teleosemantics, *in* D. Ryder, J. Kingsbury & K. Williford, eds, 'Millikan and her Critics', John Wiley & Sons, pp. 21–40.

Neander, K. (2015), 'Why I'm not a Content Pragmatist'. Unpublished paper presented at the 2015 Minds Online Conference - the Brains Blog.

Nelkin, N. (1994), 'Patterns', *Mind & Language* **9**(1), 56–87.

Newman, M. H. A. (1928), 'Mr. russell's 'causal theory of perception", *Mind* **XXXVII**(146), 137–148.

Noppeney, U., Friston, K. J. & Price, C. J. (2004), 'Degenerate neural systems sustaining cognitive functions', *Journal of Anatomy* **205**, 433–442.

O'Brien, G. (2011), 'Defending the semantic conception of computation in cognitive science', *Journal of Cognitive Science* **12**, 381–399.

O'Brien, G. & Opie, J. (2004), Notes toward a structuralist theory of mental representation, *in* H. Clapin, ed., 'Representation in Mind: New Approaches to Mental Representation', Elsevier.

O'Brien, G. & Opie, J. (2009), 'The role of representation in computation', *Cognitive Processing* **10**, 53–62.

O'Keefe, J. A. & Nadel, L. (1978), *The hippocampus as a cognitive map*, Oxford University Press.

Papineau, D. (1987), *Reality and Representation*, Blackwell Publishing.

Peacocke, C. (1994), 'Content, computation and externalism', *Mind & Language* **9**(3), 303–335.

Peacocke, C. (1999), 'Computation as involving content: a response to Egan', *Mind & Language* **14**(2), 195–202.

Perlman, M. (2000), *Conceptual Flux*, Kluwer Academic Publishers.

Piccinini, G. (2004), 'Functionalism, computationalism, and mental contents', *Canadian Journal of Philosophy* **34**(3), 375–410.

Piccinini, G. (2007*a*), 'Computational modelling vs. computational explanation: Is everything a Turing Machine, and does it matter to the philosophy of mind?', *Australasian Journal of Philosophy* **85**(1), 93–115.

Piccinini, G. (2007*b*), 'Computing mechanisms', *Philosophy of Science* **74**(4), 501–526.

Piccinini, G. (2008*a*), 'Computation without representation', *Philosophical Studies* **137**, 205–241.

Piccinini, G. (2008*b*), 'Some neural networks compute, others don't', *Neural Networks* **21**, 311–21.

Piccinini, G. (2015), *Physical Computation: a Mechanistic Account*, Oxford University Press.

Piccinini, G. & Bahar, S. (2013), 'Neural computation and the computational theory of cognition', *Cognitive Science* **34**, 453–488.

Piccinini, G. & Craver, C. (2011), 'Integrating psychology and neuroscience: functional analyses as mechanism sketches', *Synthese* **183**(3), 283–311.

Piccinini, G. & Maley, C. J. (2014), The metaphysics of mind and the multiple sources of multiple realizability, *in* M. Sprevak & J. Kallestrup, eds, 'New Waves in Philosophy of Mind', Palgrave Mcmillan.

Price, C. J. & Friston, K. J. (2005), 'Functional ontologies for cognition: the systematic definition of structure and function', *Cognitive Neuropsychology* **22**(3/4), 262–275.

Prinz, J. J. (2000), 'The duality of content', *Philosophical Studies* **100**, 1–34.

Prinz, J. J. (2002), *Furnishing the Mind: Concepts and Their Perceptual Basis*, MIT Press.

Putnam, H. (1988), *Representation and Reality*, MIT Press.

Ramsey, W. M. (2007), *Representation Reconsidered*, Cambrigde University Press.

Ramsey, W. M. (2015), 'Against representation deflation'. Presentation at the conference 'Mental Representations, the foundation of Cognitive Science?', Ruhr-Universitaet Bochum, 21-23 September 2015. Available online at: http://www.youtube.com/watch?v=mgq6uo9cDdY.

Rathkopf, C. (2013), 'Localization and intrinsic function', *Philosophy of Science* **80**, 1–21.

Reichenbach, H. (1938), *Experience and Prediction*, Chicago University Press.

Rescorla, M. (2012*a*), 'Are computational transitions sensitive to semantics?', *Australasian Journal of Philosophy* **90**(4), 703–721.

Rescorla, M. (2012*b*), 'How to integrate representation in computational modeling and why we should', *Journal of Cognitive Science* **13**, 1–38.

Rescorla, M. (2013), 'Against structuralist theories of computational implementation', *British Journal for the Philosophy of Science* **64**, 681–707.

Rescorla, M. (2014*a*), 'The causal relevance of content to computation', *Philosophy and Phenomenological Research* **LXXXVIII**(1), 173–208.

Rescorla, M. (2014*b*), 'A theory of computational implementation', *Synthese* **191**, 1277–1307.

Rupert, R. (2008), 'Causal theories of mental content', *Philosophy Compass* **3**(2), 353–80.

Ryder, D. (2004), 'SINBAD neurosemantics: A theory of mental representation', *Mind & Language* **2**, 211–240.

Ryder, D. (2009*a*), Problems of representation I: nature and role, *in* F. Garzon & J. Symons, eds, 'The

Routledge Companion to the Philosophy of Psychology', Routledge.

Ryder, D. (2009*b*), Problems of representation II: naturalising content, *in* F. Garzon & J. Symons, eds, 'The Routledge Companion to the Philosophy of Psychology', Routledge.

Scheutz, M. (1999), 'When physical systems realize functions', *Minds and Machines* **9**, 161–196.

Scheutz, M. (2001), 'Computational versus causal complexity', *Minds and Machines* **11**, 543–566.

Scheutz, M. (2012), 'What it is not to implement a computation: a critical analysis of Chalmers' notion of implementation', *Journal of Cognitive Science* **13**, 75–106.

Schweizer, P. (2014), Algorithms implemented in space and time, *in* 'Selected Papers from the 50th Anniversary Convention of the AISB', pp. 128–136.

Schweizer, P. (2016), In what sense does the brain compute?, *in* V. C. Müller, ed., 'Computing and Philosophy', Vol. 375 of *Synthese Library*, Springer, pp. 63–79.

Searle, J. R. (1980), 'Minds, brains and programs', *Behavioral and Brain Sciences* **3**(3), 417–457.

Searle, J. R. (1992), *The Rediscovery of the Mind*, MIT Press.

Shadmehr, R. & Wise, S. P. (2005), *Computational Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*, MIT Press.

Shagrir, O. (1999), 'What is computer science about?', *The Monist* **82**(1), 131–149.

Shagrir, O. (2001), 'Content, computation and externalism', *Mind* **438**, 369–400.

Shagrir, O. (2006), 'Why we view the brain as a computer', *Synthese* **153**, 393–416.

Shagrir, O. (2012*a*), 'Can a brain possess two minds?', *Journal of Cognitive Science* **13**, 145–165.

Shagrir, O. (2012*b*), 'Computation, implementation, cognition', *Minds & Machines* **22**, 137–148.

Shagrir, O. (2012*c*), 'Structural representations and the brain', *British Journal for the Philosophy of Science* **63**(3), 519–45.

Shapiro, L. A. (2000), 'Multiple realizations', *Journal of Philosophy* **97**(12), 635–654.

Shapiro, L. A. (2016), 'Mechanism or Bust? Explanation in Psychology', *British Journal for the Philosophy of Science* .

Shea, N. (2007), 'Consumers need information: supplementing teleosemantics with an input condition', *Philosophy and Phenomenological Research* **LXXV**(2), 404–435.

Shea, N. (2013*a*), Millikan's isomorphism requirement, *in* J. Kingsbury, D. Ryder & K. Williford., eds, 'Millikan and her Critics', Wiley-Blackwell.

Shea, N. (2013*b*), 'Naturalising representational content', *Philosophy Compass* **8**(5), 496–509.

Shea, N. (2014), 'Exploited isomorphism and structural representation', *Proceedings of the Aristotelian Society* **CXIV, Part 2**, 123–144.

Shine, J. M., Eisenber, I. & Poldrack, R. A. (2016), 'Computational specificity in the human brain', *Behavioral and Brain Sciences* .

Sprevak, M. (2010), 'Computation, individuation, and the received view on representation', *Studies in History and Philosophy of Science* **41**, 260–270.

Sprevak, M. (2011), 'Review: William M. Ramsey, Representation Reconsidered', *The British Journal for the Philosophy of Science* **63**(3), 669–75.

Sprevak, M. (2012), 'Three challenges to Chalmers on computational implementation', *Journal of Cognitive Science* **13**, 107–143.

Sprevak, M. (2013), 'Fictionalism about neural representations', *The Monist* **96**, 539–560.

Stich, S. P. (1983), *From Folk Psychology to Cognitive Science: The Case Against Belief*, The MIT Press.

Stich, S. P. (1992), 'What is a theory of mental representation?', *Mind* **101**(402), 243–61.

Sullivan, J. A. (2010), 'A role for representation in cognitive neurobiology', *Philosophy of Science* **77**(5), 875–887.

Suárez, M. (2003), 'Scientific representation: against similarity and isomorphism', *International Studies in the Philosophy of Science* **17**(3), 225–44.

Suárez, M. (2004), 'An inferential conception of scientific representation', *Philosophy of Science* **71**(5), 767–779.

Swoyer, C. (1991), 'Structural representation and surrogative reasoning', *Synthese* **87**, 449–508.

ter Hark, M. (2001), Wittgenstein and Dennett on patterns, *in* S. Schroeder, ed., 'Wittgenstein and

Contemporary Philosophy of Mind', Palgrave, pp. 85–103.

Tinbergen, N. (1963), 'On aims and methods of ethology', *Zeitschrift fuer Tierpsychologie* **20**, 410–433.

Tolman, E. C. (1948), 'Cognitive maps in rats and men', *Psychological Review* **55**(4), 189–208.

Usher, M. (2001), 'A statistical referential theory of content: Using information theory to account for misrepresentations', *Mind & Language* **16**(3), 311–34.

Van Gelder, T. (1995), 'What might cognition be, if not computation?', *The Journal of Philosophy* **92**(7), 345–381.

Waskan, J. A. (2006), *Models and Cognition: Prediction and Explanation in Everyday Life and in Science*, The MIT Press.

Wimsatt, W. C. (1972), 'Teleology and the logical structure of function statements', *Studies in the History and Philosophy of Science* **3**(1), 1–80.

Woodward, J. (2002), 'What is a mechanism', *Philosophy of Science* **69**, S366–S377.

Woodward, J. (2003), *Making Things Happen: A theory of causal explanation*, Oxford University Press.

Wright, L. (1973), 'Functions', *The Philosophical Review* **82**(2), 139–168.